VOL. 62, NO. 1 JANUARY 1999

THE MODERN LAW REVIEW

PUBLISHED FOR THE MODERN LAW REVIEW LIMITED
BY BLACKWELL PUBLISHERS OXFORD UK AND BOSTON, USA

THE MODERN LAW REVIEW

Volume 62

EDITORIAL BOARD

J.S. Anderson
M.R. Chesterman
Lord Dahrendorf of Clare Market
Gerald Dworkin
J.M. Evans
Anthony Giddens
Cyril Glasser
J.A G. Griffith
Judge F. Mancini
Ha[illegible] McGregor
T[illegible]
[illegible] G. Patel
Lord Wedderburn of Charlton, *Editor Emeritus*
Michael Zander

EDITORIAL COMMITTEE

Julia Black, *Production Editor*
Roger Brownsword, *Case Note Editor*
Hugh Collins, *Articles Editor*
W.R Cornish
Neil Duxbury, *Reviews Editor*
Judith Freedman
Carol Harlow
Nicola Lacey
Jenny Levin
Martin Loughlin
W.T. Murphy, *General Editor*
R.W. Rawlings
S A. Roberts
Colin Scott, *Secretary*
[illegible] C. Simpson, *Legislation and Reports Editor*
[illegible] G. Snyder
[illegible]ka Szyszczak

LP 2168

Department of Law
Acc. No. 120459
Date.....

Published for The Modern Law Review Limited by Blackwell Publishers.

© The Modern Law Review Limited 1999

ISSN 0026-7961

The Modern Law Review Ltd is incorporated in England as a company limited by guarantee under the Companies Act (Reg No 325 282) and is registered as a charity (Reg No 264072)

The Modern Law Review is published six times a year, in January, March, May, July, September and November by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main St, Malden, MA 02148, USA.

Information for contributors, subscription information and details about microfilm and advertising appear on the inside back cover.

The Modern Law Review website is at *http //www lse ac uk/depts/law/mlr.htm*

Printed and bound in Great Britain by The Alden Press, Oxford
This journal is printed on acid-free paper.

COPYRIGHT: All rights reserved Apart from fair dealing for the purposes of research or private study, or criticism or review, as permitted under the UK Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing of the Publisher, or in accordance with the terms of photocopying licences issued by organisations authorised by the Publisher to administer reprographic rights, such as the Copyright Licensing Agency and the Copyright Clearance Center.

*** FREE COURSEPACK COPYING IN UK TEACHING INSTITUTIONS DURING 1999***

For a trial period only (1 January to 31 December 1999) CLA copyright fees and permissions for the inclusion of articles, notes and reviews from any volume of Modern Law Review in study packs for use by students in UK teaching institutions have been waived. Such copies may be made free of charge during this period provided the institution does not profit from such activities.

Dept. of Law
Acc. No. LP 758
Date. 11.06.01

CONTENTS

ARTICLES

Risk and Responsibility 1
Anthony Giddens

Rights and Security in Housing: the Repossession Process in the Social Rented Sector 11
Tom Mullen, Suzie Scott, Suzanne Fitzpatrick and Robina Goodlad

Company Law and the Myth of Shareholder Ownership 32
Paddy Ireland

Universities, Defamation and the Internet 58
Douglas W. Vick, Linda Macpherson and Sarah Cooper

LEGISLATION

The Human Rights Act and Parliamentary Democracy 79
K.D. Ewing

Law and Landscape: The Legal Construction and Protection of Hedgerows 100
Jane Holder

CASES

Court-Authorised Caesarean Sections – The End of a Trend? 115
Sabine Michalowski

Pre-natal Injury and Homicide Following *Attorney-General's Reference (No 3 of 1994)* 128
Stephen Gough

Directors' 'Tortious' Liability: Contract, Tort or Company Law? 133
Ross Grantham and Charles Rickett

REVIEW ARTICLE

Equity is Dead. Long Live Equity! 140
Alison Dunn

REVIEWS

Otlowski: Voluntary Euthanasia and the Common Law 151
Rose (ed): Failure of Contracts: Contractual, Restitutionary and Proprietary Consequences 154
Kennedy: A Critique of Adjudication (fin de siècle) 156

THE
MODERN LAW REVIEW

Volume 62 **January 1999** **No 1**

Risk and Responsibility

Anthony Giddens*

I have to begin with a qualification. I am not a lawyer, and my knowledge of legal theory is at best strictly limited. So I cannot guarantee that what I have to say will even interest most of my audience, let alone prove instructive. Much of what I want to talk about concerns risk, which so far as I know does not figure prominently in legal writing. I shall say less about responsibility, which is much closer to the usual concerns of the law. I shall attempt to show, however, that the ideas of risk and responsibility are in fact closely linked.

Let me begin by posing a question. What do the following have in common: BSE; the troubles at Lloyds; the Nick Leeson affair; global warming; drinking red wine; declining sperm counts? All reflect a vast swathe of change affecting our lives today. Much of this change is bound up with the impact of science and technology on our everyday activities and on the material environment. The modern world, of course, has long been shaped by the influence of science and scientific discovery. As the pace of innovation hots up, however, new technologies penetrate more and more to the core of our lives; and more and more of what we feel and experience comes under the scientific spotlight.

The situation does not lead to increasing certainty about, or security in, the world – in some ways the opposite is true. As Karl Popper above all has shown, science does not produce proof and can never do more than approximate to truth. The founders of modern science believed it would produce knowledge built on firm foundations. Popper supposes by contrast that science is built on shifting sands. The first principle of scientific advance is that even one's most cherished theories and beliefs are always open to revision. Science is thus an inherently sceptical endeavour, involving a process of that constant revision of claims to knowledge.

The sceptical, mutable nature of science was for a long time insulated from the wider public domain – an insulation which persisted so long as science and technology were relatively restricted in their effects on everyday life. Today, we are all in regular and routine contact with these traits of scientific innovation. The consequences for health of drinking red wine, for example, were once seen by researchers as basically harmful. More recent research indicates that, taken in

* London School of Economics and Political Science.

This is the text of the twenty-sixth Chorley Lecture, delivered at the London School of Economics and Political Science on 27 May 1998 The article draws upon material from 'Risk Society: the Context of British Politics' in J. Franklin (ed), *The Politics of Risk Society* (Cambridge: Polity Press, 1998).

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

moderation, the health benefits of red wine outweigh the drawbacks. What will tomorrow's research show? Will it perhaps reveal that red wine is toxic after all?

We don't, and we can't, know – yet all of us, as consumers, have to respond in some way or another to this unstable and complex framework of scientific claims and counterclaims. Living in the UK, should one eat beef? Who can say? The health risk appears to be slight. Yet there is at least the possibility of an outbreak of BSE-related disease five, ten or twenty years from now among the human population.

We don't and can't know – the same applies to a diversity of new risk situations. Take, for instance, declining sperm counts. Some scientific studies make authoritative claims about increasing male infertility, and trace this to the action of environmental toxins. Other scientists, however, dispute the very existence of the phenomenon, let alone the explanations offered to account for it. Global warming is accepted as real by the majority of specialists in the area. Yet there is no shortage of experts who either deny that global warming exists or regard it as produced by long-term climatic fluctuations rather than by the greenhouse effect.

The Lloyds insurance market seems for the moment to have got over the disastrous financial troubles which have plagued it over the last few years. Such troubles were popularly portrayed as being bound up with class – with the complacent outlook of the 'names' and their brokers. In fact, they had their basic origin in the changing character of risk. Lloyds was hit by, among other things, findings about the toxic nature of asbestos and by a series of natural disasters – which were possibly not 'natural' at all, but influenced by global climatic change. The number of typhoons, hurricanes and other climatic disturbances happening in the world each year has climbed over the past fifteen years or so. With its massive future commitments, Lloyds – in common with other lesser insurance institutions – could be financially crippled at any time by as yet quite unforeseen negative consequences of new scientific findings or technological changes.

Simon Sebag Montefiore has written an interesting account of the adventures of Nick Leeson and Barings Bank. Sebag Montefiore suggests that there are two different ways in which what happened at Barings can be interpreted (much like the events at Lloyds). On the one hand, there is a class plus corruption explanation. According to this view, Barings Bank collapsed because it had a crusty, upper-class management at odds with the demands of a dynamic global economic order.

Sebag Montefiore casts doubt on this explanation. He argues that people working at the outer edges of the financial system, particularly in futures markets – complex markets where deals can be struck over movements in prices which have not yet, and may never, happen – are like astronauts. They have stepped outside the realm of bankers and financial experts – and they have stepped outside without a lifeline. Nick Leeson drifted away much too far from any solid ground, but most others are able to keep themselves attached to their space capsule.

Sebag Montefiore has a very arresting phrase to describe this situation. He says Nick Leeson and other people like him 'operate at the outer edge of the ordered world, on the barbaric final frontier of modern technology'. In other words, they are involved with systems which even they themselves do not understand, so dramatic is the onrush of change in the new electronic global economy. I think this is right, but the argument can be further generalised. It is not just people like Nick Leeson, not just the new financial entrepreneurs, who live at the barbaric outer edge of modern technology. *All* of us now do – and I would take this to be the

© The Modern Law Review Limited 1999

defining characteristic of what Ulrich Beck calls risk society. A risk society is a society where we increasingly live on a high technological frontier which absolutely no one completely understands and which generates a diversity of possible futures. The origins of risk society can be traced to two fundamental transformations which are affecting our lives today. Each is connected to the increasing influence of science and technology, although not wholly determined by them. The first transformation can be called *the end of nature*; and the second *the end of tradition.*

The end of nature does not mean a world in which the natural environment disappears. It means that there are now few if any aspects of the physical world untouched by human intervention. The end of nature is relatively recent. It isn't something, of course, which can be precisely dated, but we can nevertheless roughly plot when the end of nature happened. It happened when a transition came about from the sort of anxieties people used to have about nature to a new set of worries. For hundreds of years, people worried about what nature could do to us – earthquakes, floods, plagues, bad harvests and so on. At a certain point, somewhere over the past fifty years or so, we stopped worrying so much about what nature could do to us, and we started worrying more about what we have done to nature. The transition makes one major point of entry in risk society. It is a society which lives 'after nature'.

However, it is also a society which lives after tradition. To live after the end of tradition is essentially to be in a world where life is no longer lived as fate. For many people – and this is still a source of class division in modern societies – diverse aspects of life were established by tradition as fate. It was the fate of a woman to be involved in a domestic milieu for much of her life, to have children and look after the house. It was the fate of men to go out to work, to work until they retired and then – quite often soon after retirement – essentially to fade away. We no longer live our lives as fate, in a process which Ulrich Beck calls individualisation. A society which lives after nature and after tradition is really very different from the earlier form of industrial society – the basis for the development of the core intellectual traditions of Western culture.

To analyse what risk society is, one must make a series of distinctions. First of all, we must separate risk from hazard or danger. Risk is not, as such, the same as hazard or danger. A risk society is not intrinsically more dangerous or hazardous than pre-existing forms of social order. It is instructive in this context to trace out the origins of the term 'risk'. Life in the Middle Ages was hazardous; but there was no notion of risk and there doesn't seem in fact to be a notion of risk in any traditional culture. The reason for this is that dangers are experienced as given. Either they come from God, or they come simply from a world which one takes for granted. The idea of risk is bound up with the aspiration to control and particularly with the idea of controlling the future.

The observation is important. The idea of 'risk society' might suggest a world which has become more hazardous, but this is not necessarily so. Rather, it is a society increasingly preoccupied with the future (and also with safety), which generates the notion of risk. The idea of risk, interestingly, was first used by Western explorers when they ventured into new waters in their travels across the world. From exploring geographical space, it came to be transferred to the exploration of time. The word refers to a world which we are both exploring and seeking to normalise and control. Essentially, 'risk' always has a negative connotation, since it refers to the chance of avoiding an unwanted outcome. But it can quite often be seen in a positive light, in terms of the taking of bold initiatives

© The Modern Law Review Limited 1999

in the face of a problematic future. Successful risk-takers, whether in exploration, in business or in mountaineering, are widely admired.

We should distinguish risk from hazard, but we must also make a distinction between two kinds of risk. The first two hundred years of the existence of industrial society were dominated by what one might call *external risk*. External risk, expressed in down-to-earth terms, is risk of events that may strike individuals unexpectedly (from the outside, as it were) but that happen regularly enough and often enough in a whole population of people to be broadly predictable, and so insurable. There are two kinds of insurance associated with the rise of industrial society: the private insurance company, and public insurance, which is the predominant concern of the welfare state.

The welfare state became the left's project in the post-1945 period – it became seen above all as a means of achieving social justice and income redistribution. By and large, however, it did not originate as such. It developed as a security state, a way of protecting against risk, where collective rather than private insurance was necessary. Like early forms of private insurance, it was built on the presumption of external risk. External risk can be fairly well calculated – one can draw up actuarial tables and decide on that basis how to insure people. Sickness, disablement, unemployment were treated by the welfare state as 'accidents of fate', against which insurance should be collectively provided.

A world which lives after nature and after the end of tradition is one marked by a transition from external to what I call *manufactured risk*. Manufactured risk is risk created by the very progression of human development, especially by the progression of science and technology. Manufactured risk refers to new risk environments for which history provides us with very little previous experience. We often don't really know what the risks are, let alone how to calculate them accurately in terms of probability tables.

Manufactured risk is expanding in most dimensions of human life. It is associated with a side of science and technology which the early theorists of industrial society by and large did not foresee. Science and technology create as many uncertainties as they dispel – and these uncertainties cannot be 'solved' in any simple way by yet further scientific advance. Manufactured uncertainty intrudes directly into personal and social life – it isn't confined to more collective settings of risk. In a world where one can no longer simply rely on tradition to establish what to do in a given range of contexts, people have to take a more active and risk-infused orientation to their relationships and involvements.

The rise of risk society has several interesting consequences – which should concern anyone who has taken an interest in the BSE debate in Britain and continental Europe, or in fact in any of the episodes I mentioned at the beginning of this discussion.

As manufactured risk expands – or, if you like, as we live more and more in a risk society in Ulrich Beck's terms – there is a new riskiness to risk. In a social order in which new technologies are chronically affecting our lives, and an almost endless revision of taken-for-granted ways of doing things ensues, the future becomes ever more absorbing, but at the same time opaque. There are few direct lines to it, only a plurality of 'future scenarios'.

We recently saw the tenth anniversary of the nuclear disaster at the Chernobyl plant. No one knows whether it is hundreds – or millions – of people who have been affected by the Chernobyl fall-out. The long-term effects will in any case be difficult to chart, because if they exist they are likely to be diffuse. We are altering the environment, and the patterns of life we follow, almost constantly. Even many

 © The Modern Law Review Limited 1999

apparently benign habits or innovations could turn sour – just as, conversely, risk can often be overestimated. Take the example of smoking. Smoking was encouraged by doctors up to some thirty or so years ago as a means of relaxation. No one knew the time bomb which the practice of smoking was stirring up. The BSE episode might have an opposite outcome. Perhaps it will turn out that humans are not affected. It is characteristic of the new types of risk that it is even disputed whether they exist at all.

In risk society there is a new moral climate of politics, one marked by a push-and-pull between accusations of scaremongering on the one hand and of cover-ups on the other. A good deal of political decision-making is now about managing risks – risks which do not originate in the political sphere, yet have to be politically managed. If anyone – government official, scientific expert or lay person – takes any given risk seriously, he or she must proclaim it. It must be widely publicised because people must be persuaded that the risk is real – a fuss must be made about it. However, if a fuss is indeed created and the risk turns out to be minimal, those involved will be accused of scaremongering.

Suppose on the other hand that the authorities decide that the risk is not very great, as the British government did in the case of BSE. In this case, the government says: we've got the backing of scientists here; there isn't much risk, we can go on as we did before. Yet if things turn out otherwise, then of course they will be accused of a cover-up.

Paradoxically, scaremongering may be necessary to reduce risks we face – yet if it is 'successful' in this sense, it appears as just that, scaremongering. The case of AIDS is an example. Suppose governments and experts make great public play with the risks associated with unsafe sex, to get people to change their sexual behaviour, and AIDS does not spread nearly as much as originally predicted. The response is likely to be: why were you scaring everyone like that? This sort of political dilemma becomes routine in risk society, but there is no easily available way of confronting it. For as I mentioned earlier, even whether there are any risks at all is likely to be controversial. We just cannot know beforehand when we are actually 'scaremongering' and when we are not.

The emergence of a risk society is not wholly about the avoidance of hazards, for reasons also given previously. Risk society, looked at positively, is one in which there is an expansion of choice. Now obviously choice is differentially distributed according to class and income. As nature and tradition release their hold, for instance, some otherwise infertile women can pay to have children through the use of new reproductive technologies, whereas others cannot. We know that in detraditionalised social settings some women live in poverty after divorce, whereas others achieve a more rewarding life than they could have done before. Technological innovation usually expands the domain of choice; as does the disappearance of tradition. As customary ways of doing things become problematic, people must choose in many areas which used to be governed by taken-for-granted norms. Eating is an example: there are no traditional diets any more.

The advent of risk society has strong implications for rethinking the political agenda in this country and elsewhere. The emergence of manufactured risk presumes a new politics because it presumes a reorientation of values and the strategies relevant to pursuing them. There is no risk which can even be described without reference to a value. That value may be simply the preservation of human life, although it is usually more complex. When there is a clash of the different types of risk, there is a clash of values and a directly political set of questions.

© The Modern Law Review Limited 1999

Modernisation as, for example, Tony Blair uses the term, means bringing Britain up to date. Tony Blair has been the archetypal moderniser within the Labour Party; but more fundamentally, he wants to modernise British institutions – modernisation carrying the connotation in this country that Britain lags behind other industrial societies in various key respects. Now this is a bit like the first explanation that Sebag Montefiore mentions for the collapse of Barings Bank – crusty old institutions which have lost their relevance to the modern world.

That there is something in the project of modernisation, thus understood, can be seen by anyone who sets foot in the House of Lords. In risk society, however, modernisation means something different. Risk society is industrial society which has come up against its own limitations, where those limitations take the form of manufactured risk. Modernisation in this sense, cannot simply be 'more of the same'.

We should distinguish here between simple and reflexive modernisation. Simple modernisation is old-type unilinear modernisation; reflexive modernisation, by contrast, implies coming to terms with the limits and contradictions of the modern order. These are obvious in new domains of politics associated with various sorts of social movements. They are obvious in motorway protests, in animal rights demonstrations and in many food scares. Second-phase modernisation – modernisation as reflexive modernisation – will not look like first-phase modernisation. There is an opportunity, I think, for political debate in this country to leap ahead of many other European countries in this respect and I would like to see this happen. Reflexive modernisation, like risk more generally, is by no means wholly a negative prospect and offers many possibilities for positive political engagement.

Our relationship to science and technology today is different from that characteristic of early industrial society. In Western society, for some two centuries, science functioned as a sort of tradition. Scientific knowledge was supposed to overcome tradition but actually became a taken-for-granted authority in its own right. It was something which most people respected, but was external to their lives. Lay people 'took' opinions from the experts. The more science and technology intrude into our lives, the less this external perspective holds. Most of us – including government authorities and politicians – have, and have to have, a much more dialogic or engaged relationship with science and technology than used to be the case. We cannot simply 'accept' the findings which scientists produce, if only because scientists so frequently disagree with one another, particularly in situations of manufactured risk. And everyone now recognizes the essentially sceptical character of science described earlier. Whenever someone decides what to eat, what to have for breakfast, whether to drink decaffeinated or ordinary coffee, that person takes a decision in the context of conflicting, changeable scientific and technological information.

There is no way out of this situation – we are all caught up in it, even if we choose to proceed 'as if in ignorance'. Politics must give some institutional form to this dialogical engagement, because at the moment it concerns only special interest groups, who mostly struggle outside the main political domain. We do not currently possess institutions which allow us to monitor technological change. We might have prevented the BSE debacle if a public dialogue had already been established about technological change and its problematic consequences. Enoch Powell apparently remarked that nothing affects our lives as much as technological change and he was right – yet such change is completely outside the democratic system. More public means of engaging with science and technology wouldn't do

© The Modern Law Review Limited 1999

away with the quandary of scaremongering versus cover-ups, but it might allow us to mute some of its more damaging consequences.

These considerations are relevant to rethinking the welfare state. The welfare state was founded against the backdrop of a society where nature was still nature and tradition was still tradition. This is obvious, for example, in the gender provisions in the post-1945 welfare state, which completely presumed the continuity of the 'traditional family'. It is obvious in terms of the growth of the National Health Service, which was set up as a response mechanism to illness understood as external risk.

In a world of more active engagement with health, with the body, with marriage, with gender, with work – in an era of manufactured risk – the welfare state cannot continue on in the form in which it developed in the post-1945 settlement. The crisis of the welfare state is not purely fiscal, it is a crisis of risk management in a society dominated by a new type of risk.

These observations are relevant to class division. J. K. Galbraith's so-called 'culture of contentment' was a bit of a shooting star – there is no culture of contentment. One reason why many middle-class and professional groups have opted out of public welfare schemes is bound up with a certain attitude towards risk management. In risk society, the middle classes detach themselves from public provision and in a certain sense they are right to do so because that provision was geared to a different interpretation and situation of risk. When people have a more active orientation to their lives, they also have to have a more active orientation to risk management, so it is not surprising that those who can afford it tend to opt out of existing welfare systems.

Ecological questions precisely reflect a world living after nature and after tradition. Many forms of lifestyle politics develop which have no precedent in the earlier type of industrial society. Protesters some while ago made a great deal of fuss about veal calves being transported to the continent in constrained and artificial conditions. Their critics called them sentimental. Yet in the light of the experience of BSE, everyone can see that this wasn't just sentiment. The protests reflected a sense of what can happen when the industrial production of food becomes distanced from nature – or what used to be nature. A moral commitment to animal rights is, in a certain sense, a hard-edged politics – after all, even measured in narrow economic terms, the BSE crisis has been a disaster. Calculations put the cost of the British economy at £6 billion or perhaps even more.

Risk society is not the same as postmodernism. Postmodern interpretations see politics as at an end – political power simply loses its significance with the passing of modernity. Yet modernity does not disappear with the arrival of manufactured risk; rather modernisation, which continues, takes on new meanings and subtleties. Reflexive modernisaton presumes and generates a politics. That politics cannot unfold completely outside the parliamentary domain. Social movements and special interest groups cannot supply what parliamentary politics offers – the means of reconciling different interests with one another, and also a balance of different risks in relation to one another. The issues I have discussed demand to be brought more directly into the political arena. A party able to address them cogently would be in a prime position in the political encounters that will unfold over the coming few years.

Risk is always related to security and safety. It is also always connected to responsibility. It isn't surprising therefore that as we move towards a world dominated by manufactured rather than external uncertainty, there is a renewed

© The Modern Law Review Limited 1999

discussion of the nature of responsibility. Widespread use of 'responsibility' is also quite recent. Although the word 'responsible' is much older, 'responsibility' only seems to have come into the English language in the late eighteenth century. It is again a notion associated originally with the rise of modernity. As it is used today, 'responsibility' is an interestingly ambiguous or multi-layered term. In one sense, someone who is responsible for an event can be said to be the author of that event. This is the original sense of 'responsible', which links it with causality or agency. Another meaning of responsibility is where we speak of someone being responsible if he or she acts in an ethical or accountable manner. Responsibility also however means obligation, or liability, and this is the most interesting sense to counterpose with risk.

The relation between risk and responsibility can be easily stated, at least on an abstract level. Risks only exist when there are decisions to be taken, for reasons given earlier. The idea of responsibility also presumes decisions. What brings into play the notion of responsibility is that someone takes a decision having discernable consequences.

The transition from external to manufactured risk is bringing about a crisis of responsibility, because the connections between risk, responsibility and decisions alter. This is a crisis of responsibility with negative and positive features, roughly corresponding to the negative and positive aspects of risk. Given the inherently ambiguous nature of most situations of manufactured risk, and the inherent reflexivity of these situations, responsibility can neither easily be attributed nor assumed. This applies both where responsibility means limiting risk (as in ecological risks, or health risks) and where risk is an energising principle (financial markets).

Several consequences follow:

1. The emergence of what Beck calls 'organised irresponsibility'. By this he means that there are a diversity of humanly created risks for which people and organisations are certainly 'responsible' in a sense that they are its authors but where no one is held specifically accountable. Various questions then come to the fore. Who is to determine how harmful products are, what side effects are produced by them, and what level of risk is acceptable? How can 'sufficient proof' be determined in a world full of contested knowledge claims and probabilities? If there are damages to be paid, or reparations made, who is to decide about compensation and appropriate forms for future control or regulation?

   Much of the 'social interrogation' of risk and responsibility takes place through the prism of external risk and simple modernisation. This is true, for example, of anyone who expects an actuary to predict risk, and therefore assess responsibility, on the basis of past trends; or of anyone who supposes that one can simply turn to experts to provide solutions. Coping with situations of organised irresponsibility is likely to become more and more important in the fields of law, insurance and politics, but this won't be easy to do precisely because of the rather imponderable character of most circumstances of manufactured risk. The dilemma of scaremongering versus cover-ups is a direct indication of the deep seated nature of the problems involved here.
2. Some say that the most effective way to cope with the rise of manufactured risk is to limit responsibility by adopting the 'precautionary principle'. The notion of the precautionary principle seems to have first emerged in

 © The Modern Law Review Limited 1999

Germany in the 1980s, in the context of the ecological debates that were carried on there. At its simplest, it proposes that action on environmental issues (and by inference other forms of risk) should be taken even though there is scientific uncertainty about them. Thus in the 1980s, in several Continental countries, programmes were initiated to counter acid rain, whereas in Britain lack of conclusive evidence was used to justify inactivity on this and other pollution problems too. Yet the precautionary principle isn't always helpful or even applicable as a means of coping with problems of responsibility. The precept of 'staying close to nature', or of limiting innovation rather than embracing it, can't always apply. The reason is that the balance of benefits and dangers from scientific and technological advance, and other forms of social change, is imponderable. We may need quite often to be bold rather than cautious in supporting scientific innovation or other forms of change.

This having been said, variations on the precautionary principle can nevertheless be a significant way of reintroducing responsibility. One variant of the principle, for example, is that firms producing goods should think through the whole product cycle before those goods are released onto the market or relevant technical processes utilised. Thus in the Brent Spar episode, the company putting up the oil platform in the first place had not adequately thought through to the final point of effective and reasonably safe disposal.

3. Situations of manufactured risk shift the relation between collective and individual responsibility in many risk situations. Although in many circumstances individuals cannot be held culpable, this is not the same as non-culpability in conditions of organised irresponsibility. In the latter case, this results from viewing responsibilities through the lenses of external or passive risk. Consider, for instance, health risks. Many people get ill through no fault of their own. But a large proportion of illnesses are related both to lifestyle practises and to wider conditions of the 'created environment'. It doesn't make any sense to suppose that liability in these circumstances can remain wholly with the collectivity, whether this be government or an insurance company. The active assumption of responsibility, as in attempts to reduce levels of smoking, becomes part of the very definition of risk situations and therefore the attribution of responsibility. Something quite similar applies to our responsibilities towards future generations. When most risk was external, such responsibility was relatively limited: nature was largely intact. Our responsibilities to future generations now are thoroughly infused with decisions we have to take resulting from our transformation of nature.
4. These considerations are relevant to one of the major political issues of our times, the future of the welfare state. The history of the welfare state in all countries is a tangled one. The welfare state emerged in some part as a means of holding back the aspirations of the poor and of controlling them – it had some of its roots in the political right. In recent years, however, as described earlier, the left has appropriated the welfare state as its own project. The debate around the welfare state has therefore concentrated to a considerable degree upon its role in limiting or reducing inequality. But the welfare state is more correctly seen as a form of collective risk management. The idea that the welfare state should be understood as a 'safety' or 'provident' state has been raised most forcefully in the writings

© The Modern Law Review Limited 1999

of the French thinker Francois Ewald. The welfare state is tied into the basic suppositions of modernity – that security comes from the ever more effective control by human beings of their material and social environments.

The crisis of the welfare state is usually represented as a fiscal one. If the welfare state is in trouble, it is because people won't pay the taxes needed to fund welfare systems properly. There is some validity to this, but it is more illuminating to see the crisis of the welfare state as a crisis of risk management. The welfare state was built up on the presumption of external or passive risk. If you become unemployed, fall ill, become disabled or lose your home, the welfare state will step in to protect you. Welfare systems must now confront large areas of manufactured risk, shifting the relation between risk and responsibility. It isn't surprising that there is now a great deal of talk about the need to connect rights with responsibilities. Unconditional rights might seem appropriate when individuals bear no responsibility for the risks they face, but such is not the case in situations of manufactured risk.

5. Where a society hasn't got effective means of dealing with organised irresponsibility, the result isn't always that no one is held culpable. On the contrary, the price of manufactured uncertainty is probably closely associated with the emergence of the 'litiginous' society. Where a common 'contract of responsibility' has broken down, culpability can appear everywhere. Here indemnity has effectively been separated from causality. I might be held responsible, for example, if someone is hurt through slipping on my garden path.
6. The theme of responsibility has to be integrated with a concern for the two sides of risk. The negative and positive sides of risk are still often discussed as though they were separate from one another. This translates into a division between two large bodies of literature. It is a remarkable fact that most of those who write about environmental risk make no reference at all to the literature on financial or entrepreneurial risk, or vice versa. Two of the most influential books to have been written about risk over the past ten years, for example, are *Risk Society* by Ulrich Beck and *Against the Gods* written by Peter Bernstein. Yet these books make no reference at all to one another.

UNIVERSITY OF ★ DEPARTMENT OF LAW NOT TO BE REMOVED

The fact that risk is often a positive or energising phenomenon is relevant to most of the situations of risk and responsibility discussed above, not just to economic risk. Thus to create a more effective welfare state, it is important that in some situations people are psychologically and materially able to take risks albeit in a 'responsible' way. It isn't a good outcome for the individual or the wider society where a person is stuck on benefits or unwilling to take the risk of plunging into the labour market. The same applies to someone caught up in a dysfunctional or violent relationship. Risk is not only closely associated with responsibility, but also with initiative and the exploration of new horizons – something which takes us back to our starting point when the notion was first developed in post-medieval Europe.

The themes of responsibility and culpability have obviously always been of interest to lawyers. I hope I have given at least some indication as to why legal theorists and practitioners should also concern themselves with the idea and reality of risk as well.

© The Modern Law Review Limited 1999

LAW CAMPUS LIBRARY ACC. NO. UNIVERSITY OF GLASGOW
LP2168
Department of Law Library Acc. No. J20459 Date.....

Rights and Security in Housing: the Repossession Process in the Social Rented Sector

Tom Mullen, Julie Scott, Suzanne Fitzpatrick and Robina Goodlad***

Rented housing in Britain is divided between a private sector in which houses are let for profit, and a social rented sector in which houses have traditionally been let as a service to the community. The social rented sector is itself divided between two main types of provider: local authorities (council housing), and housing associations, although account must be taken of other providers, particularly in Scotland, where Scottish Homes, a quango, is a substantial provider of social housing. Until recently, the new towns were substantial providers of rented housing, but the new town development corporations have now all been wound up and their stock transferred to other landlords. In this article, the expressions 'social rented sector' and 'social housing' refer to housing let by the public sector bodies just referred to and by housing associations.

For most of this century the nature of the provider has determined the legal tenure on which rented housing has been held. Thus, the Rent Acts applied to private sector lets, but were not applied to public sector tenants. Between 1980 and 1989 council tenants and housing association tenants shared a common tenure regime, the secure tenancy. However, when in 1989, a new form of tenure – the assured tenancy – was introduced for the private sector,[1] it was also applied to new housing association lets, although local authorities and other public sector bodies continued to offer secure tenancies. As a consequence, since 1989, there have been two different legal tenures within the social rented sector: the secure tenancy and the assured tenancy. The two legal regimes afford substantially different rights to tenants. In the light of these changes to tenure regimes, this article discusses the importance of legal guarantees of security of tenure in the social rented sector, and compares the way in which the process of recovery of possession has been operating in the two forms of tenure. This is a two-fold comparison. As the rights of existing housing association tenants were preserved in 1989, housing associations continued to have many secure tenants, and so it is possible to compare the respective positions of secure and assured tenants both (a) as between local authorities and housing associations, and (b) within the housing association sector.

The research on which this article is based shows that there are substantial differences in the outcomes of the repossession process for different groups of tenants. The questions these outcomes pose include whether differences in the legal rights of secure and assured tenants are capable of explaining the outcomes of the repossession process, and what other factors might be influencing outcomes.

* School of Law, University of Glasgow
** Centre for Housing Research and Urban Studies, University of Glasgow

1 By the Housing (Scotland) Act 1988. Assured tenancies were introduced for England and Wales by the Housing Act 1988.

LP758 11.06.01

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

These questions are themselves relevant to broader themes about the role of law in a regulated housing system and the appropriate legal framework for securing the social rights of citizenship.[2]

The focus of this article is the social rented sector in Scotland. But although Scotland has its own regime of landlord and tenant law, the broad thrust of housing policy and the aims of housing legislation have been similar throughout Britain in recent years.[3] Many of the insights derived from the research may, therefore, be transferable, provided allowance is made for legal and contextual differences between the social rented sectors in Scotland and elsewhere in Britain. Important differences between Scottish and English law are highlighted in the text or footnotes.

The research

The research which provides the data for this article[4] was funded by Scottish Homes, a public agency which took over the functions of the Housing Corporation in Scotland, including funding and regulating housing associations, after the Housing (Scotland) Act 1988.[5] The overall aim of the research was to provide a picture of the extent to which the position of tenants in the social rented sector had been affected by the introduction of assured tenancies under Part II of the Housing (Scotland) Act 1988 ('the 1988 Act'), with particular, but not exclusive, emphasis on issues of security of tenure. The research had four elements: first, an analysis of the tenancy agreements used by social landlords in Scotland; second, a questionnaire survey on landlords' repossession policies, practices and activity; third, an analysis of court records of repossession cases brought by housing associations in four Sheriff courts; and, fourth, case studies of housing management in nineteen social landlords. The fieldwork and data analysis were carried out between January and August 1996. The questionnaire data relate to the financial years 1994–95 and 1995–96. The court records are those for the calendar year 1995. The research was carried out immediately before the local government reorganisation of April 1996 and, therefore, all local authority data refer to pre-reorganisation authorities.

The housing policy context

The principal concern of housing policy in relation to rented housing in the 1980s and 1990s has been to increase the role of the market. The public sector has been reduced by a combination of measures including granting public sector tenants the

2 See for example, M Partington, 'Citizenship and Housing' in R. Blackburn (ed) *Rights of Citizenship* (London: Mansell, 1993)

3 Whether the important changes made by the Housing Act 1996 which apply only to England and Wales mean that we are embarking on a period of divergent development remains to be seen The principal changes in the 1996 Act – new rules on allocation of housing, reduced duties to the homeless, introductory tenancies, and the raft of changes related to anti-social behaviour – were not applied to Scotland, although some relatively minor aspects of the first and last mentioned have already led to amendments to Scottish legislation, including those in the Crime and Disorder Act 1998.

4 T. Mullen, S. Scott, S. Fitzpatrick and R Goodlad, *Tenancy Rights and Repossession Rates in Theory and Practice* (Edinburgh: Scottish Homes, 1997).

5 Unlike the Housing Corporation, Scottish Homes is also a significant social landlord, but it intends to divest itself of its housing stock over the next few years

 © The Modern Law Review Limited 1999

right to buy their homes (Tenants' Rights Etc (Scotland) Act 1980) and restrictions on capital expenditure which largely prevent the construction of new council houses. The 1987 White Paper, *Housing: The Government's Proposals for Scotland*,[6] and the 1988 Act marked a further development of this approach. They confirmed the shift away from public sector bodies (local authorities, new towns, Scottish Homes) as providers of new social housing. Instead housing associations were to be preferred providers. Local authorities were to have a role as 'enablers' and all public sector landlords were to be encouraged to transfer their stock to new landlords including housing associations.

In order to hasten the run-down of the public sector, Part III of the 1988 Act introduced a new mechanism (somewhat misleadingly labelled 'tenants choice') to permit public sector tenants (tenants of local authorities, new towns and Scottish Homes) to transfer to new landlords provided the new landlord was either outside the public sector or was Scottish Homes. However, most stock transfers took place not by this route but, in order to secure investment, by 'voluntary transfer' under pre-existing powers which were amended to introduce requirements for consulting tenants and ensuring majority consent.[7] Despite the hopes expressed in the White Paper, the private sector has been almost completely uninvolved in this process. All transfers from local authorities and Scottish Homes have been to housing associations, except for two transfers to non-profit local housing companies set up specifically to take over Scottish Homes stock. The rate of transfer has been determined largely by the availability of public funds channelled through Scottish Homes.

The 1987 White Paper also proposed a new approach to the private sector: decontrolling rents and reducing security of tenure in order to 'revitalise' it.[8] The necessary legal changes were introduced by Part II of the 1988 Act. From the beginning of 1989, new lets in the private sector were to be on assured tenancies which could be let at market rents and for which there was only a vestigial form of rent control. The tenant under an assured tenancy still had security of tenure, but of a lesser degree than that which had been enjoyed by regulated or secure tenants under the Rent Acts. Landlords also had the option of letting on *short assured* tenancies[9] in which recovery of possession at the end of the lease was guaranteed.

The 1988 Act had important consequences for housing associations. Before 1980, housing association tenancies appeared to be part of the private sector, at least for legal purposes. Then the Tenants' Rights Etc (Scotland) Act 1980 appeared to treat housing association tenants as part of the public sector in that they held their houses on secure tenancies and, therefore, enjoyed substantially the same rights as public sector tenants, including the right to buy the house they lived in.[10] The one major departure from the public sector model was that rents continued to

6 Cm 242 (1987) Similar proposals for England and Wales were made in *Housing: The Government's Proposals*, Cm 214 (1987)

7 The provisions applying to local authorities are the Housing (Scotland) Act 1987, ss 12, 12A, 81B and Sched 6A. Other provisions apply to Scottish Homes and new towns In England and Wales, local authority housing may be voluntarily transferred under the Housing Act 1985 s 32. The Housing Act 1988 introduced two further procedures for transfer. The 'Tenants Choice' provisions in Part IV of that Act have since been repealed by the Housing Act 1996, but Part III providing for transfer of housing to Housing Action Trusts established by the Secretary of State remains in force.

8 Cm 242, para 1.19. For England and Wales, see *Housing: The Government's Proposals*, Cm 214 (1987)

9 The English equivalent is the assured shorthold tenancy.

10 The right to buy was extended to housing association tenants by the Housing (Scotland) Act 1986.

© The Modern Law Review Limited 1999

be fixed by the rent officer. However, the 1988 Act applied the new private sector model to housing associations. From the beginning of 1989, new housing association tenants were given assured rather than secure tenancies and were subjected to the same decontrolled rent system and reduced security of tenure as tenants of private sector landlords.

The tenure changes were associated with the changes in the financial regime. Housing associations had been funded largely through housing association grant (HAG), introduced by the Housing Act 1974. This covered the bulk of the capital cost of development leaving little debt to be recovered from current rents. The 1987 White Paper promised a reduction in the level of HAG with an increasing proportion of the cost of development to be met by private sector finance. Implicit in the White Paper was the assumption that private sector funders would only be attracted if housing associations switched to the new private sector model of the assured tenancy.

However, the switch to assured tenancies might not have been equally attractive to tenants. The form of tenure became an issue in stock transfers as tenants were made aware that transferring out of the public sector to a housing association would mean a change in their status from secure to assured tenants.[11] This would entail the loss of statutory rights, including reduction in security of tenure, although the right to buy is retained in most cases of voluntary transfer.[12]

Housing associations were reluctant to identify themselves with private landlords and were lukewarm about the tenure changes.[13] The movement's representative body, the Scottish Federation of Housing Associations (SFHA), published and recommended to its members a Model Assured Tenancy Agreement (MATA)[14] which was intended to replicate the statutory rights of secure tenants by contract. In apparent contradiction to the 1987 White Paper stance, the SFHA's approach has been endorsed by Scottish Homes whose guidance assumes that housing associations will issue the MATA to their tenants rather than take full advantage of the opportunity to extend the landlord's freedom of action afforded by the 1988 Act.[15]

It is difficult to gauge the precise impact of differences between secure and assured tenancies on stock transfers given the limited empirical evidence available on tenants' perceptions and attitudes. However, ballots in the Scottish new towns have indicated that, where the council is a choice, tenants tend to vote overwhelmingly for this option, and tenant surveys have indicated that rent levels and security were the most important factors informing that choice. In Glenrothes and East Kilbride, where housing associations were on the list of potential new landlords, tenants seemed unconvinced that the improved tenancy rights offered in housing association leases would be binding.[16] Although our research sheds no light on the issue of tenants' perceptions, it is relevant to the issue of how important in practice are the undoubted legal differences between tenures.

---

11 S. Scott, *1988 and all That: The Changing Role of Housing Associations in Scotland*, Occasional Paper 31 (Glasgow: CHRUS, University of Glasgow, 1997).

12 The right to buy was preserved for tenants who became assured tenants as a consequence of a voluntary transfer. See the Housing (Scotland) Act 1987, s 81A.

13 Scottish Federation of Housing Associations, *Response to Consultation Paper* (unpublished, 1987).

14 SFHA, (Edinburgh: 1990).

15 Scottish Homes and SFHA, *Raising Standards in Housing* (Edinburgh. 2nd ed, 1992).

16 R. Goodlad and S. Scott, 'Housing and the Scottish New Towns: A Case Study of Policy Termination and Quasi-markets' (1996) 33 *Urban Studies* 317.

© The Modern Law Review Limited 1999

Security of tenure: an outline of the legislation[17]

Before looking at the research findings it is essential to appreciate the main differences in statutory rights between secure and assured tenants with regard to security of tenure. Secure tenancies are now governed by Part III of the Housing (Scotland) Act 1987 ('the 1987 Act')[18] and assured tenancies by Part II of the 1988 Act.[19] In both tenures a tenant may be evicted only on one or more of the grounds specified in the legislation whatever it may say in the contract.[20] In both tenures there is a distinction between grounds on which recovery of possession is allowed only if the court considers it reasonable and grounds on which recovery of possession is mandatory. We will not enumerate all the differences between the secure and assured grounds, but concentrate on those which are likely to seem most significant to tenants. By far the commonest reason for eviction is rent arrears and it is here that most important differences between the tenures are to be found. In a secure tenancy there is a single rent related ground – ground 1.[21] The court may award recovery of possession where some rent is in arrears but it must also consider it reasonable to do so. By contrast, in an assured tenancy there are three rent related grounds for recovery of possession. Ground 12 is the nearest equivalent to the secure tenancy ground 1: it permits recovery of possession in any case where there are rent arrears provided the sheriff considers it reasonable. In addition, ground 11 allows for possession where the tenant has persistently delayed in paying rent, *even if there are no current arrears*, but again only at the sheriff's discretion. However, under ground 8 recovery of possession is mandatory where an assured tenant owes at least three months rent,[22] both when the notice of proceedings[23] is served and at the date of the court hearing.

Ground 2 also represents a significant difference for assured tenants. It allows for recovery of possession where the landlord has defaulted on a loan and the heritable creditor requires vacant possession in order to realise the security. There is no remotely equivalent ground for secure tenancies.

In addition to the differences in the grounds for possession, there is also a technical point to be made about the different structures of secure and assured tenancies which is of direct relevance to the SFHA's aim of replicating secure tenants' rights by contract. The secure tenancy is a statutory status which can be ended only in the manner prescribed in the 1987 Act, but there is also a contract between the landlord and the tenant. The tenant enjoys the benefits of both the contract and the statutory status from the beginning of the tenancy until the time the tenancy is terminated in accordance with the 1987 Act. By contrast, in an assured tenancy under the 1988 Act, security of tenure is guaranteed in a different way. There is a clear distinction between the initial contractual tenancy agreed between the parties and the statutory assured tenancy which may follow it if the contract is validly terminated. In an assured tenancy, the contract can be ended

17 For a general account of the law, see C. Himsworth, *Housing Law in Scotland* (Edinburgh: Butterworths, 4th ed, 1994); P. Robson, *Residential Tenancies* (Edinburgh: W. Green, 1994).

18 For secure tenancies in England and Wales, see Part IV of the Housing Act 1985.

19 For assured tenancies in England and Wales, see Part I of the Housing Act 1988.

20 1987 Act, s 47 and Sched 3; 1988 Act s 18 and Sched 5.

21 Ground 1 also permits recovery of possession for a breach of any other obligation of the tenancy

22 The amount of arrears under the equivalent ground in England and Wales was reduced to two months by the Housing Act 1996, s 101.

23 The notice of proceedings is intended to warn the tenant of impending court proceedings and gives the tenant either two weeks or two months warning. See 1988 Act, s 19 and SI 1988/2109. A similar notice is required in secure tenancies. See 1987 Act, s 47 and SI 1980/1389.

© The Modern Law Review Limited 1999

simply by giving notice to quit at the appropriate time in accordance with the terms of the contract. But, if the contract is validly terminated, it is followed by a statutory assured tenancy, under which the tenant continues to have a tenancy of the house so long as s/he retains possession of it.[24] This is in marked contrast to the position taken in English law where there is no division between contractual and 'statutory' tenants, and a notice to quit is neither required nor effective for the purpose of terminating an assured tenancy. The Scottish provisions do, however, parallel those applying to regulated tenancies in England and Wales under the Rent Act 1977.[25]

In both the contractual and the statutory assured tenancy the landlord's right to recover possession of the house is restricted to the grounds specified in Schedule 5 to the 1988 Act. The terms of any statutory assured tenancy are the same as those of the prior contractual tenancy, except that any terms relating to termination of the tenancy and rent increases do not carry over.[26] As we shall see, the retention of the distinction between the contractual and the statutory tenancy leads to complications in the drafting of housing association leases which would not occur in England and Wales.

It is clear that the legislation gives assured tenants a lesser degree of security than secure tenants because eviction of assured tenants is permitted on additional grounds (2, 8 and 11) to those available in relation to secure tenants. However, a social landlord which was reluctant to take full advantage of the greater management freedom permitted by the 1988 Act could use the lease to put its assured tenants in approximately the same position as secure tenants. The landlord could undertake to evict only on grounds which would be available in a secure tenancy, for example, in the case of rent arrears, only on ground 12. For this strategy to be effective the contract must be drafted in a particular way.

The 1988 Act states that only certain grounds for recovery are available in a *contractual* assured tenancy, and only if they are written into the tenancy agreement.[27] They include the heritable creditor and all three rent grounds. But these restrictions on the grounds available do not apply in statutory assured tenancies. Moreover, as noted above, those terms of the contract relating to termination do not carry over to a statutory assured tenancy. Thus all the grounds may be made available by simply terminating the tenancy. Since tenancy agreements in the social rented sector are typically of short duration, often monthly, an assured tenancy contract which leaves out the 'extra' grounds (2, 8, and 11) may present little obstacle to swift recovery of possession under those grounds.[28] Accordingly, the only sure way to confer on assured tenants an equivalent degree of protection from eviction to that enjoyed by secure tenants, is to draft the contract so that the right to recover possession is restricted to the equivalent grounds *and* it expressly states that the landlord will not seek to terminate the contract by notice to quit except on one of the equivalent grounds.

---

24 1988 Act, s 16.

25 It also parallels the provisions applying to regulated tenancies in Scotland – see the Rent (Scotland)Act 1984.

26 1988 Act, s 16 There is an exception whereby a rent increase clause drafted in a particular form will continue to have effect.

27 1988 Act, s 18.

28 Where the lease contains a positive undertaking to use only stated grounds, but no restriction on the right to issue a notice to quit, there might be scope for a personal bar argument (equivalent to estoppel) if the landlord did attempt to use grounds not in the original agreement to evict a statutory assured tenant.

© The Modern Law Review Limited 1999

Security of tenure: the role of tenancy agreements

We discussed in the previous section the possibility of using the lease itself to give assured tenants an equivalent degree of security of tenure to that enjoyed by secure tenants. Indeed this was what the SFHA tried to do by publishing the MATA, the aim of which was said to be to 'build in contractual rights for housing association tenants which are as close as possible to existing secure tenants rights'.[29] The guidance accompanying the MATA itself goes slightly further in saying 'the SFHA has sought to replicate the rights of secure tenants within this contractual assured tenancy agreement'.

The general intention is clear, but although in many respects the MATA replicates secure tenants' rights, it falls short of granting assured tenants the same level of protection in relation to repossession as secure tenants. The first point to make is that it gives landlords a degree of choice because some clauses (or sub-clauses) are designated as 'core' and others as 'optional'. The guidance accompanying the MATA states that core clauses must be included by all housing associations using the MATA as the basis for their tenancy agreements whereas optional clauses 'should be included by housing associations as they are examples of good practice'. Therefore, the degree to which any particular agreement replicates secure tenants' rights depends in part on the precise combination of 'core' and 'optional' clauses which a housing association adopts from the MATA.

The second point relates to the grounds which are included in the MATA either as 'core' or 'optional' sub-clauses. Ground 2 from the 1988 Act (loan default by landlord) is omitted because 'it is regarded as unduly onerous to the tenant'[30] so in that respect the gap between secure and assured rights is narrowed. But, all three rent-related grounds are included. Ground 11 (persistent delay in paying rent) is given as a core clause and ground 8 (mandatory recovery for three months rent arrears) as an optional clause. Furthermore, it expressly confers on the landlord the right to issue a notice to quit, without stating any qualification or restriction. Thus, even where a housing association using the MATA omits on its own initiative grounds 2, 8, and 11 from the list of grounds for recovery, these grounds can become available simply by terminating the contract. There would be no great delay in gaining access to these grounds because the MATA is a monthly tenancy.

For the housing association movement as a whole to replicate secure tenants' rights would require associations to take a conscious decision to omit the MATA clauses relating to grounds 8 and 11, to avoid any reference to ground 2, and to insert in their leases a restriction on the right to terminate by notice to quit along the lines described in the preceding section. There is neither reference to the latter point in the guidance nor draft wording for an appropriate notice to quit clause.

In fact, we found that nearly all Scottish housing associations adopted some variant of the MATA termination clause in their assured tenancy agreements. Our analysis indicated that just over three-quarters (77 per cent – 105 cases) of Scottish housing associations[31] included ground 8 in their contractual assured tenancy and

29 SFHA, *The Housing (Scotland) Act 1988: A Practical Guide* (Edinburgh: 1989).

30 Guidance accompanying the MATA.

31 The study excluded housing associations which did not offer assured tenancies. these were mainly fully mutual co-operatives and specialist providers. Tenancy agreements were received from 137 out of the 162 (85 per cent) housing associations included in the survey

© The Modern Law Review Limited 1999

almost nine tenths (88 per cent – 121 cases) made provision for Ground 11. Only 13 associations (less than 10 per cent) stated only ground 12 as a rent arrears ground in their tenancy agreement, and thus approximated to secure tenants' rights. Ground 2 was included in the tenancy agreements of one third of associations (34 per cent – 46 cases) despite its exclusion from the MATA and the specific guidance on that point.[32] However, ground 2 cannot be used to recover possession unless the landlord has given written notice to the tenant that it might be used on or before creation of the tenancy.[33] It was not possible to assess from analysis of the tenancy agreements what proportion of associations served the notice required to make this ground effective.[34]

No assured tenancy agreement attempted to restrict the landlord's right to issue notice to quit to circumstances equivalent to the grounds for repossession available under a secure tenancy. This is unsurprising given that the issue is not raised in the guidance to the MATA, and associations may not have had this technical point drawn to their attention by their legal advisers. Two associations did, however, attempt to limit the grounds upon which they could seek repossession of the resulting *statutory assured* tenancy, and specifically excluded grounds 2, 8 and 11. Strictly speaking these contractual provisions do not carry over to the statutory tenancy, but they might become legally effective indirectly, as the undertaking might lead to an argument in any court case that the association were personally barred from using the excluded grounds.[35] Most of the associations in the case studies indicated that they would only issue a notice to quit when seeking possession, and would only do so on the grounds they used for contractual tenants. Thus in practice it appears that associations may not use the device of the notice to quit to circumvent self-imposed restrictions in their tenancy agreements.

Most housing associations, therefore, adopted a tenancy agreement whose terms in relation to termination were consistent with the MATA, but given the terms of the MATA, this has meant that the housing association movement has failed to achieve its aim of replicating secure tenants' rights for assured tenants. If this is simply due to unreflective adoption of the MATA then the remedy would be to revise the MATA on the assumption that housing associations would adopt the revised version. In fact, the SFHA did revise the MATA in the light of the research findings, and published the new version in July 1997.[36] The accompanying guidance states that the SFHA gave serious consideration to including only ground 12. However, both grounds 8 and 11 remain in the text. In the case of ground 8, housing associations are 'strongly urged to use it only as a last resort, and even then as sparingly as possible'. The relevant clause also contains an undertaking not to issue a notice to quit except where the landlord intends to take court action for recovery of possession. Therefore, even the new MATA does not fully replicate secure tenants' rights. As we shall see below, a significant number of housing associations are keen to retain the extra grounds for recovery of possession, so it is unlikely that this aim will be achieved in the near future.

32 There may not have been an intention to include ground 2 in all of these cases. In most of them, ground 2 was included by implication rather than expressly – rather than enumerating specific grounds, the leases included a general reference to all the grounds in the 1988 Act.

33 1988 Act, s 18, and Sched 5, para 2. The sheriff has a discretion to dispense with this requirement.

34 Where the inclusion of ground 2 was inadvertent the failure to serve the notice would not be surprising. See n 32 above.

35 See n 28 above.

36 SFHA, Model Assured Tenancy Agreement (Glasgow: SFHA, revised ed, 1997).

© The Modern Law Review Limited 1999

Security of tenure: repossession policies and practices

The next question we examined was whether there were material differences in outcomes for secure and assured tenants, as well as a difference in legal rights: were assured tenants more likely to lose their homes? Here we had three sources of information: (i) a questionnaire sent to all social landlords in the study, (ii) the analysis of court records of housing association eviction cases, and (iii) the case studies. We were able to compare our results with earlier findings of the Accounts Commission, the baseline study of public sector landlords carried out in 1993, and a 1990 study of housing associations.[37] Although data on all eviction actions were collected, the vast majority related to rent arrears.[38] We compiled statistics relating to each stage of the legal process:

(i) issuing a notice of proceedings;
(ii) issuing a summons to a court;
(iii) obtaining a decree for recovery of possession; and
(iv) enforcing a decree for possession.

Cases do not necessarily proceed through all the stages. A case may be settled at any stage and this is frequently done where the parties make an arrangement for the repayment of arrears. Even where a court hearing takes place, repossession may be refused for a variety of reasons if the case is contested, on the basis for example that there is an error in procedure or pleadings, that the relevant ground for recovery has not been established, or that it is not reasonable to evict. The statistics for each stage of the process are summarised in Table 1.

Table 1: Comparison of legal activity 1995–96, rates per 1,000 tenancies

	Public Sector	Housing Associations		
		Secure	Assured	all HA
Notices seeking possession	159	*	*	28
Court action	25	3.7	7.1	6.3
Decrees	13.8	1.4	4.5	3.4
Evictions	1.01	0.7	1.6	1.48

* Information on the number of notices of proceedings served by housing associations was not collected separately for secure and assured tenants.

Source: 1996 Questionnaire Survey

Notices of proceedings

The first formal step in the legal process is the service on the tenant of a notice of proceedings, the purpose of which is to warn the tenant that the landlord may be

37 Accounts Commission, *Tenants' Rent Arrears – A Problem?* (Edinburgh: 1995); D. Clapham, K. Kintrea, J. Malcolm, H. Parkey, and S. Scott, *A Baseline Study of Housing Management in Scotland* (Edinburgh: HMSO, 1995); A. Kearns and J. Malcolm, *Housing Management among Scottish Housing Associations: Research Report No 30* (Edinburgh: Scottish Homes, 1994).

38 In our study 99 per cent of local authority evictions were for rent arrears. For housing associations the figure were 93 per cent in 1994–95 and 87 per cent in 1995–96. These figures are consistent with the earlier studies.

© The Modern Law Review Limited 1999

about to begin possession proceedings.[39] As Table 1 shows, public sector landlords served notices at a far greater rate than housing associations. The average in 1995–96 was 159 per 1,000 tenancies. Rates appear to have risen since the 1993 baseline study which found that the average rate of issuing notices was 128 per 1,000. On average, housing associations served 28 notices per 1,000 tenancies in 1995–96. The average rate is very similar to the rate of service in the 1990 study (27 per 1,000), but as the sector has grown substantially the actual number of notices served has almost doubled over the period from 1,160 to 2,163 notices. In both sectors, the averages conceal wide variations in activity. The rate of service of notices in the local authority sector varied between four per 1,000 and 457 per 1,000, and in the housing association sector between none per 1,000 and 195 per 1,000.

Court action

The next stage is raising a court action. Public sector landlords were also much more likely than housing associations to take tenants to court. The average rate for court action was, at 25 cases per 1,000 tenancies, about four times the overall rate of 6.3 cases for housing associations. For both sectors, this is an increase, in the public sector from 20.1 per 1,000 tenancies in 1993 and, amongst housing associations, from 4.1 per 1,000 in 1990. However, within the housing association sector there was a clear difference between assured and secure tenants. Assured tenants were taken to court at an average rate of 7.1 per 1,000, whereas the average rate for secure tenants was only 3.7 per 1,000. Again, there were great variations within each sector.

Decrees for possession

In 1995–96, public sector landlords obtained decrees for possession at an average rate of 13.8 per 1,000 tenancies, an increase on the average rate of 7.4 per 1,000 tenancies found in the 1993 baseline study. In 1995–96, housing associations obtained decrees for possession at the much lower rate of 3.4 per 1,000 tenancies. However, within this sector, the average rate for assured tenants was much higher at 4.5 per 1,000 tenants than that for secure tenants (1.4 per 1,000). Once again rates vary widely within sectors.

Enforcement of decree

Decrees for possession are frequently not enforced and the tenant is often allowed to remain in the property. The rate of actual eviction has increased for all housing association tenants from 1.08 per 1,000 tenancies in the 1990 study to 1.48 per 1,000 in 1995–96 (the total number of evictions being 94). Again, there is a significant difference between housing association secure and assured tenants. The eviction rate for assured was 1.6 per 1,000 tenancies, and only 0.7 for secure tenants.

By comparison, the overall eviction rate for public sector landlords in 1995–96 was only 1.01 per 1,000 tenancies. Although this is double the rate found in 1993 (average 0.5 per 1,000), the proportion of tenants actually evicted is still considerably lower than in the housing association sector. Eviction rates in the

39 1987 Act, s 47, 1988 Act, s 19.

© The Modern Law Review Limited 1999

public sector are only 68 per cent of those in the housing association sector. It is also apparent that given the much greater frequency of obtaining a decree for possession in the public sector, public sector landlords *enforce* a far smaller proportion of the decrees they obtain than do housing associations. On average 7.3 per cent of public sector decrees are followed by eviction, as compared to 43.5 per cent of housing association decrees.

Again, there are wide variations within sectors, and these are perhaps more significant in the housing association sector. All public sector landlords evicted some tenants, but about two thirds (67 per cent) of housing associations did not evict anyone in 1995–96. In general these are associations with no arrears or very low arrears levels. If we exclude from the analysis those associations with arrears of less than one per cent of gross rent receivable, the rate for all housing association evictions rises to 1.67 per 1,000. However, the figure for secure tenant evictions rises only slightly to 0.73 per 1,000 and that for assured tenants rises substantially to 2.09 per 1,000. Thus, where arrears are higher and legal action more frequent, assured tenants seem even more likely to lose their homes.

It is clear that public sector landlords serve notices of proceedings, take court action and obtain decrees at much higher rates than housing associations, but the proportion of tenants actually evicted is less in the public sector. So, in those cases in which they do take legal action, housing associations are more likely to pursue the legal process to its limits. In addition to this, within the housing association sector there are significant differences between secure and assured tenants. Assured tenants are much more likely to face court action and eviction than secure tenants, and in fact secure tenants of housing associations are even less likely to be evicted than secure tenants in the public sector. The trend is to increased rates of eviction in both sectors.

Explaining differential repossession rates

So far we have established that housing association assured tenants tend to have less security of tenure in law than secure tenants even after taking account of contractual adjustments. Our data also indicate two sets of differences in the outcomes of the eviction process. The first is that between public sector and housing association tenants; the second is that within the housing association sector between its secure and assured tenants. We now need to consider the possible causes.

It would be rash to assume that these differences in outcomes are explained by differences in the legal rights of the groups of tenants concerned since a variety of factors might help to explain such outcomes. The research casts light on the role of some of the possible factors but the research could not examine all plausible causative factors directly. Our initial hypothesis here is that three sets of factors might be particularly important. These are differences in the legal rights of tenants, differences in landlords' approaches to arrears management and differences in the characteristics of tenants. The research data relate primarily to the legal and managerial factors, rather than to tenant characteristics. However, other published data enable us to draw tentative conclusions about the role of tenant characteristics. These three types of factor all relate to the immediate context of the eviction process, but are not intended to exclude consideration of the wider context which includes central government policy, funding arrangements and controls on the two sectors. This wider context is also considered in relation to arrears management.

© The Modern Law Review Limited 1999

Legal factors: nature of the tenure

If legal rights were the major determinant of outcomes (the legal rights hypothesis) we would expect that to be reflected in higher repossession rates for assured tenants as compared to both housing association secure tenants and public sector tenants (all of whom are secure). Table 1 shows that eviction rates are indeed higher for assured tenants than for both groups of secure tenants. However, closer examination of the figures suggests that neither of these differences in eviction rates supports the legal rights hypothesis. It is clear that eviction rates are lower for housing association secure tenants than for public sector secure tenants (0.7 per 1,000 compared to 1.01 per 1,000) which immediately suggests that there are important differences between the sectors other than legal rights. Also, looking at the legal process as a whole, rather than final outcomes, casts a different light on the comparison between public sector tenants and housing association assured tenants. Public sector tenants are much more likely than housing association tenants to have proceedings raised against them, and the higher rates of eviction for housing association assured tenants are largely a function of the fact that housing associations convert decrees for possession into evictions at a much higher rate than do public sector landlords. This is significant because the enforcement of a decree for possession is at the discretion of the landlord and does not necessarily reflect legal differences in the tenancy regimes.

The stage at which legal differences might be expected to make most difference is the decision whether to grant a decree for possession, since that is a decision for the court. It is then that the differences in the grounds for possession ought to bite, because the court has to decide whether the legal requirements for awarding possession are satisfied. If, therefore, we look at the proportion of actions raised which result in decrees, we find that the proportion of court actions which result in a decree for possession is higher for housing association assured tenants than for public sector tenants (63 per cent compared to 55 per cent). However, the proportion of actions in which possession is granted is lowest of all for housing association secure tenants (38 per cent) again suggesting that factors other than legal rights are important.

If the hypothesis is revised so that legal differences are merely one of several variables, but not necessarily the most important, then it might be possible to explain the results. For example, if there were substantial differences in approaches to arrears management between the two sectors such that public sector landlords resorted to court action more readily than housing associations, that could explain why the greater likelihood of legal action being taken in the public sector appears to co-exist with more favourable rights for tenants in that sector. The legal factors might then be sufficient to explain why there are more evictions of assured than secure tenants *within* the housing association sector.

To support the revised version of the legal rights hypothesis we would have to find some evidence suggesting that higher eviction rates for housing associations' assured tenants are linked to their less advantageous legal position. The additional grounds for possession were certainly relied upon in practice. Although ground 2 (landlord's loan default), does not appear to have been used, grounds 8 (three months rent arrears) and 11 (persistent delay in paying rent) were frequently used. Of the two, ground 8, as the only rent ground on which possession is mandatory, ought to be the more important. The questionnaires showed that more than six out of 10 associations (62 per cent) said that they rarely or never used this ground, but

 © The Modern Law Review Limited 1999

a substantial minority, approximately three out of ten (31 per cent), said that they would always or usually use it.

So, it is clear that the additional rent grounds are widely used in practice. The next question is whether there is any evidence that the use of the additional rent grounds increases the likelihood that a repossession order will be granted. If it did we would expect to find two things. First, that housing associations' assured tenants were more likely than their secure tenants to have an eviction decree awarded against them; second, that in those cases brought against assured tenants decrees were more likely to be awarded in cases brought on grounds 8 and/or 11 than in cases brought on ground 12 alone.

The questionnaires did not ask on which precise ground proceedings were raised so we cannot compare the outcomes of cases raised on different grounds, but the results of the analysis of court records provided useful information on this point. We attempted to analyse all housing association repossession actions in four relatively busy sheriff courts in 1995. We identified 321 housing association cases. However, often the specific grounds on which the action was brought were not apparent from the records so we had to ask the associations concerned to supply details of the grounds relied upon, and were, therefore, only able to identify specific grounds in 221 cases in all. The figures relating to specific grounds in the remainder of this section (including those in Table 2) are derived from these cases. They must, therefore, be treated with some caution.

Almost all (96 per cent) cases brought by housing associations against secure tenants were brought on ground 1 in the 1987 Act and were related to rent arrears rather than the other limb of ground 1 (breach of tenancy). Similarly, 93 per cent of

Table 2: Grounds for possession, by outcome

	Outcome					
	Eviction Decree	Debt Only Decree	Expenses Only	Dismissed	Sisted[40]	Total No
Secure						
Ground 1	12	0	3	5	26	46
Ground 7[41]	1	0	0	0	0	1
Other	0	0	0	1	0	1
Assured						
Ground 8	12	1	0	4	6	23
Ground 11	2	0	0	1	1	4
Ground 12	42	4	2	7	31	86
Ground 8 & others	20	0	0	4	7	31
Grounds 11 & 12	10	0	0	0	7	17
Rent + other	1	0	0	0	1	2
Other	8	0	0	2	0	10
Total	108	5	5	24	79	221
Percentage	49	2	2	11	36	100

Source: Analysis of court data

40 Where a case is 'sisted' it is in effect adjourned indefinitely. The case will not be brought back to court unless and until one of the parties takes the initiative to recall it. Cases are usually sisted to arrive at a settlement or to monitor a repayment arrangement. Often such cases are ultimately resolved without further court process which is why it appears so frequently as a final outcome.

41 Ground 7 permits eviction of a secure tenant where a nuisance or annoyance has been caused.

© The Modern Law Review Limited 1999

the assured tenancy cases were based on one or more of the rent grounds in the 1988 Act. Ground 12 (the equivalent of secure ground 1) was used on its own in half (49 per cent) of cases. Ground 8 (the mandatory ground) was used alone or in combination with other grounds in a third (32 per cent) of cases. Ground 11 (persistent delay) was used alone or in combination with ground 12 in a further 12 per cent of cases. Therefore, repossession actions were brought on the two rent grounds not available in secure tenancies in 44 per cent of cases.

The results summarised in Table 2 indicate that over half of assured cases (54 per cent – 95 cases) ended in decree for possession compared with only a quarter (27 per cent – 13 cases) of secure cases. When we compare cases brought on the two rent arrears grounds which are directly equivalent (ground 1 for secure and ground 12 for assured tenants) we find that the proportion of cases involving assured tenants resulting in decree was almost double that for secure cases (49 per cent compared to 26 per cent). So it appears that housing association secure tenants are far less likely to have possession awarded against them than assured tenants even when the grounds for possession are legally equivalent: the particular rent related grounds on which action is brought appears to make little difference.

Looking at assured tenancy cases only, we find that decrees were awarded against assured tenants in 59 per cent of cases in which ground 8 alone or in combination had been used (32 out of 54 cases)[42] and 57 per cent of cases in which ground 11 was used alone or in combination (12 out of 21). Where ground 12 alone was used 49 per cent of cases ended in decrees (42 out of 86). There is, therefore a clear but not a strong correlation between the use of the additional possession grounds and outcomes in assured tenancies. But, when the comparison between secure and assured tenants is added in, it is clear that there is at most limited statistical support for the proposition that the additional rent grounds have made it more likely that assured tenants will have decrees taken against them. The wide gap between secure and assured decree rates, even where equivalent grounds are used, remains to be explained. Whilst it is true that the gap is certainly greater if secure rent arrears cases are compared to assured cases brought under ground 8 and 11 rather than assured cases generally, this may simply reflect the fact that these cases would be more likely to lead to the award of a decree whatever grounds were used.

Both the questionnaire and court records suggested that there were differences of views amongst housing associations on the appropriateness of using the extra rent grounds. The case studies provided further evidence of this split. Around a quarter of case study housing associations did not include ground 8 in the tenancy agreement and a further quarter included it in the agreement, but had made a policy decision not to use it. The remaining half of the case study associations did use ground 8 and offered a variety of reasons for this practice. Some thought it a clearer and fairer method than the use of a discretionary ground. A commonly held view was that eviction proceedings were only begun when all else had failed, and in circumstances in which the tenant could not reasonably complain about being evicted. To quote one housing officer:

> If someone is in three or more months arrears then the association's view is that we have by that time used every device available to us to encourage them to pay their rent. That device [ground 8] is only used where a tenant simply ignores arrangements and will not pay their rent.

42 Court records almost invariably disclose only that decree was granted, and not the specific grounds relied on by the sheriff in reaching his decision.

© The Modern Law Review Limited 1999

We must consider the possibility that the additional grounds really are having an impact upon outcomes, but that this is being masked by some combination of off-setting factors. There are a number of factors which could affect the outcome of eviction actions including the levels of attendance by and representation of tenants. However, such factors can only be masking the impact of differences in legal rights if they are affecting eviction cases unevenly according to tenure type, for example, if representation and/or attendance rates were different for secure and assured tenants, or if sheriffs had different attitudes to different types of tenant and landlord. The research did not directly address these issues. As a result, we cannot rule out this possibility. However, it does seem very difficult to imagine factors which can explain the results found whilst also being generally plausible. For example, if representation and/or attendance had a major effect on outcomes – a frequent observation in socio-legal research[43] – we would expect a smaller gap between secure and assured eviction rates than we found, and this would still not explain why the particular grounds on which action is brought makes so little difference to the gap between secure and assured eviction rates. Similarly, although we know that judicial attitudes towards different categories of litigants may sometimes be important,[44] it is not clear what set of judicial attitudes would explain our results. Nor does previous research on possession proceedings suggest that factors such as these impact differently according to tenure type, although it must be acknowledged that it was not an issue in that research.[45]

To summarise, our results suggest that legal differences explain at most a small part of the divergence in repossession rates between secure and assured tenants. We need to examine closely other factors, in particular arrears management styles and the characteristics of tenants.

Management factors: differences in landlords' approaches to rent arrears

The differing approaches of landlords to rent arrears management may have their origins in the separate financial and institutional context within which the two social housing sectors operate or, more simply, in their differing administrative practices. Previous research[46] has shown that such differences exist, for example some authorities pursued rent arrears quickly while in others it was given a very low priority. These differences appeared to be a question of practice as well as policy.

Social landlords are subject to a variety of pressures which might affect the way they manage housing, including financial pressures, their legal and social obligations and existing systems for ensuring accountability to government and the community, but our discussion of these pressures is necessarily tentative. As regards financial pressures, both sectors have been affected by cuts in public expenditure which have had the effect of pushing rents up in real terms but neither sector as a whole seems to be under objectively greater pressure from mounting arrears: average rent arrears are declining in both sectors and stood at 3.7 per cent

43 See, for example, H. Genn and Y. Genn, *Representation at Tribunals* (London Lord Chancellor's Dept, 1989).

44 J. Nixon *et al*, *Housing Cases in County Courts* (Bristol. The Policy Press, 1996).

45 For other recent research on the eviction process see, S. Mason and M. Rose-Troup, *Rent Arrears Actions in the Sheriff Courts* (Edinburgh. Shelter, 1995) and J Nixon *et al*, *Housing Cases in County Courts*, n 44 above.

46 See references at n 37 above.

© The Modern Law Review Limited 1999

in housing associations at the end of 1995, and 3.5 per cent in local authorities.[47] More significant may be the pattern of capital spending which has generally favoured housing associations in the late 1980s and early 1990s, possibly with some impact on housing management capacity and performance. Housing management is better resourced in the housing association sector, which has traditionally enjoyed higher staffing levels and lower stock/staff ratios. Kearns and Malcolm found an average of 107 properties per member of staff in housing associations in 1990[48] whereas the baseline study of public landlords found an average of 274 properties per member of staff in 1992–93.[49] Size may also be an issue.[50] The average association in the study had only 512 tenancies, whereas only 12 per cent of local authorities had less than 1,000 properties and 17 per cent had over 15,000 properties.

The general arrangements for central government supervision also differ, but both sectors have been under pressure to improve the quality of housing management. The arrangements for performance scrutiny seem to be more detailed and intrusive for housing associations and, on the face of it, funding is more closely related to performance for them. Local authorities are subject to the *Tenants Charter for Scotland*,[51] and are required to produce performance indicators by the Accounts Commission.[52] From 1991 they were preparing for compulsory competitive tendering of housing management,[53] although the CCT regime is being replaced by 'best value'. Housing associations are subject to supervision and regulation by Scottish Homes which has tried to establish a 'performance culture'.[54] Associations are required to complete an annual return based on detailed performance indicators and are graded by Scottish Homes on each aspect of performance. Funding for development has also been explicitly linked to performance.[55]

The approach of public sector landlords to the repossession process may be influenced by their distinctive role. Housing associations do not share the same form of direct political accountability of local authorities, their legal responsibility towards homeless persons, or their wider strategic housing responsibilities. Also, local authorities may perceive themselves as a landlord of last resort for their area. For these reasons they may see eviction as a less realistic or less acceptable option and thus be inclined to give tenants more leeway. Again, however, there was little direct evidence to support this in the research.

The roles of officers, management committees and councillors may be significant in explaining differences between different landlords. Our study found a much higher level of involvement of housing association committee members than local authority elected members in decisions about legal action but also found that eviction rates appeared to be lower where the management committee were

47 A. Kearns and J. Malcolm, *Housing Management among Scottish Housing Associations. Research Report No 30* (Edinburgh: Scottish Homes, 1994); Scottish Office, *Statistical Bulletin· Housing Series HSG/1996/5* (Edinburgh: HMSO, 1996).

48 *ibid.*

49 See n 37 above.

50 See D Clapham and K Kintrea, *Housing Co-operatives in Britain· Achievements and Prospects* (Harlow: Longman, 1992).

51 Scottish Office, 1991.

52 Local Government Finance Act 1992, Part I

53 Scottish Office, *Competing for Quality in Housing* (Edinburgh: 1991)

54 Scottish Homes, *Scottish Homes Future Relationship with Housing Associations* (Edinburgh: Scottish Homes, 1994).

55 *ibid.*

© The Modern Law Review Limited 1999

involved in decisions about individual cases. Whereas 67 per cent of associations had no evictions in 1995–96, the proportion having no evictions was much higher (72 per cent) where committees were involved in decisions than where they were not (52 per cent).[56]

The 'good practice' literature in rent arrears management[57] suggests that landlords should make 'settling in' visits to new tenants, carry out checks to assess entitlement to housing and other benefits[58] in order to maximise tenants' income, take early action when tenants go into arrears and monitor performance closely. Studies in England have suggested that housing associations are more likely to follow such guidance than local authorities. For example, one study found that approximately four out of ten (42 per cent) of housing associations carried out debt counselling compared with only two out of ten (21 per cent) local authority landlords.[59] Another study found that housing associations were more likely than local authorities to visit tenants as part of their early action to recover arrears.[60] In response to our questionnaire, 79 per cent of public authorities said that they always or usually took legal action to recover rent arrears, compared to only 48 per cent of housing associations.

The case studies suggested that housing associations spent more time and effort on the pre-litigation stages of arrears than public sector landlords. They were, therefore, more likely to reach satisfactory arrangements for repayment of arrears with tenants without beginning the legal process. In practice, housing associations put a heavy premium on early intervention and a 'personal approach'. For example, one association officer said: 'I see two or three tenants a week... they pay a few pounds and it gets them back into a routine.' In contrast, in three of the four case study public sector landlords, the computer generated notices automatically at a set number of weeks, although in one of these staff had a discretion on whether they were sent: 'I know that there's a view that if you serve too many notices they become like toilet paper. But the computer recommends that you serve it and we use our judgement about whether it's appropriate.'[61]

In many public sector landlords, there was little sense that decrees were likely to lead to ejection of tenants, and the summons to court was in effect a slightly stronger warning letter. In contrast, housing associations tended to enrol cases in court only when they 'really meant it' and therefore more often pursued the legal process through to eviction. They felt justified in taking this harder line once legal action had commenced because tenants were only taken to court once all else had failed. The characteristics of those evicted were often summed up by officials and committee members as 'can pay – won't pay'. One of the case study authorities

56 This is consistent with the findings of B. Gray *et al*, *Rent arrears in Local Authorities and Housing Associations in England* (London: HMSO, 1994)

57 See, for example, Accounts Commission, *Tenants' Rent Arrears – A problem? Volume 1* (Edinburgh: 1991); Chartered Institute of Housing in Scotland, *Housing Management Standards Manual*, (Edinburgh: 1996); Scottish Homes and Scottish Federation of Housing Associations, *Raising Standards in Housing* (Edinburgh: 2nd ed, 1996).

58 Incorrect assessment of housing benefit is a potentially significant cause of arrears in the non-local authority sector, especially given the availability of the mandatory ground. The case studies suggested that incorrect assessment of housing benefit was a frequently encountered problem, but that housing associations were generally successful at sorting out such difficulties.

59 University of Glasgow, Centre for Housing Research, *The Nature and Effectiveness of Housing Management in England* (London: HMSO, 1989).

60 B. Gray *et al*, n 56 above.

61 Research carried out before the introduction of secure tenancies suggested that the routine use of notices to quit diminished their impact. See D Wilkinson, *Rent Arrears in Public Authority Housing in Scotland* (Edinburgh. HMSO, 1980); See also Accounts Commission, n 57 above.

© The Modern Law Review Limited 1999

was about to begin a pilot scheme under which their approach to court action would be closer to that of housing associations: 'We will give them lots and lots of chances, but when we finally take the decision to lodge we will use the decree as soon as we get it.'

Thus, the apparently greater firmness of housing associations in enforcing decrees may merely reflect the fact that they have invested more time and energy in the pre-litigation stage of arrears recovery. Thus, it is more likely that they have 'weeded out' cases where recovery of arrears was possible. Officials and committee members would also be likely to have a stronger perception than those of public sector landlords that they had done all they could short of legal action. There are, therefore, good grounds for supposing that the distinct arrears recovery practices of the two sectors contribute substantially to explaining the striking contrast between rates of legal action (higher in the public sector) and rates of actual eviction (lower in the public sector) across the two sectors. But general differences of approach to arrears management cannot explain the enormous differences in outcomes between secure and assured tenants within the housing association sector.

The case study associations were quite firm that they treated their secure and assured tenants in exactly the same way, so far as the statutory framework would allow, and we uncovered no evidence to the contrary. Thus neither differences in legal rights, nor differences in approaches to arrears management, appear fully to explain the differences in the outcomes of the repossession process. This suggests we should examine whether the characteristics of the tenants themselves are having an impact.

The characteristics of the tenants

The questionnaire returns indicated that assured tenants are more likely to have arrears, and to have serious arrears, than secure tenants of housing associations. Only 56 per cent of respondents were able to give a breakdown of rent arrears between the different types of tenancy. However, of those who could, 19 per cent had no arrears due from their secure tenants, whereas 11 per cent had no arrears attributable to their assured tenants. It also appeared that assured tenants were more likely to have substantial arrears. Here, over three-quarters (78 per cent) of respondents could give a breakdown of those who owed more than three months arrears between tenure types. Of these associations, over half (53 per cent) said that none of their secure tenants owed more than three months rent. In contrast only a quarter (26 per cent) of these associations said that none of their assured tenants owed more than three months rent.

Why are assured tenants more likely to have substantial rent arrears? Previous studies have found that households with children and single people under retirement age are far more likely to have arrears than pensioner households or childless couples.[62] We can relate this to information about letting patterns. The Scottish Continuous Recording system (SCORE) provides information on the household characteristics of housing association lets. Almost all lets in the last few years have been on assured tenancies. The data also show that the proportion of lets to older people has been falling steadily, from 30 per cent of those housed in 1992–

62 *ibid* and S. Scott and K Kintrea, *Tenants Rent Arrears: A Problem? Volume 2* (Edinburgh: Accounts Commission, 1992).

© The Modern Law Review Limited 1999

93 to 26 per cent in 1995–96. In contrast, the proportion of lets to single people under retirement age and families has risen. Single adults secured a third (34 per cent) of the lets in 1995–96 while families with children accounted for over a quarter of the lets (27 per cent).[63]

Some of the case study associations thought that the 'younger generation' were less concerned about debt than older generations of tenants. To quote one housing officer: 'The younger ones are not so concerned about the arrears. The older ones were brought up with the idea that if you don't have the money, you don't buy it.' This view is reinforced by the example of a case study organisation which specialised in housing for the elderly where interviewees indicated that both secure and assured tenants had low arrears, and generally special needs associations figure prominently amongst those with very low arrears.

This all supports the hypothesis that assured housing association tenants are increasingly likely to be in groups which are prone to rent arrears. Secure tenants are inevitably more likely to be in older age groups and to have been tenants for longer than assured tenants because of the length of time they have been in the sector. These differences in tenant profile are the result of a combination of the legal limitations on the creation of new secure tenancies after 1988, and the changing role of housing associations which has led to them taking over much local authority stock.

Differences in tenant profiles may also help to explain part of the difference in outcomes between the public and housing association sectors, in particular the fact that public sector secure tenants appear simultaneously less at risk of eviction than housing association assured tenants, but more at risk than housing association secure tenants. Data about the differences between groups of tenants can be gleaned from the 1991 Census which provides details of the overall profile of households in Scotland. There are also two large scale surveys of tenants: a study of housing association tenants carried out in 1991,[64] and the baseline study of public sector landlords.[65] Taken together these show that housing associations house a higher proportion of pensioners (38 per cent of households) than local authorities (32 per cent). Housing associations also house a higher proportion of single adults (19 per cent) than local authorities (10 per cent). Finally the proportion of families with children living in the housing association sector (including single parent families) is low (26 per cent) in comparison with the general population (34 per cent), and very much lower than in the local authority sector (40 per cent). Given that public sector landlords house a higher proportion of tenants in one of the groups most prone to rent arrears (families with children) and housing associations house a higher proportion of tenants in one of the groups least prone to rent arrears (pensioners), it is not surprising that local authority tenants are more likely to have arrears than housing association secure tenants, and this may contribute to higher rates of legal action and eviction. Therefore, there are grounds for concluding that differences in tenant profiles may help to explain differences in outcomes for tenants both within the housing association sector and between the two sectors.

63 Scottish Federation of Housing Associations, *SCORE: New Housing Association Tenants in Scotland: The SCORE Annual Digest 1995–96*, (Edinburgh: 1996).

64 A. Kearns, D. Redmond, and J. Malcolm, *The Ins and Outs of Renting: A study of the socio-economic circumstances of Scottish housing association tenants* (1993).

65 n 37 above.

© The Modern Law Review Limited 1999

Conclusion

There are clear differences in the statutory rights of secure and assured tenants. The housing association movement has attempted to close the gap by means of contractual arrangements but to date these have not succeeded. We can, therefore say that, as a matter of law, nearly all housing association assured tenants have significantly less protection against the repossession of their homes than secure tenants. The revised MATA, although more effective for its purpose than the original version, will not necessarily succeed in closing the gap if housing associations do not wish to relinquish using the mandatory ground for eviction in rent arrears cases. Since we also found that housing association assured tenants are more likely to be evicted than both local authority and housing association secure tenants, it is tempting to assume that this is due to their weaker legal rights. However, our research indicates that specifically legal factors almost certainly cannot account for the great difference in outcomes for tenants either across the two sectors or within the housing association sector. These differences provide at most a small part of the explanation of the greater likelihood of assured tenants losing their homes.

We suggest that the greater part of the differences in outcomes between the sectors can plausibly be explained by a combination of the distinct nature of housing associations' and public sector landlords' policies and practices for recovery of rent arrears, and by differences in the profile of tenants which lead to assured tenants generally having higher levels of rent arrears. However, differences in management policies and practices do not seem to be able to explain differences in outcomes between secure and assured tenants within the housing association sector. Here the most significant factor appears to be the higher levels of arrears found amongst assured tenants, again probably related to tenant profiles: assured tenants are more likely to be in the groups prone to rent arrears.

The precise reasons for the differences in approaches to arrears management remain elusive. The more intensive approach to the pre-litigation stages of arrears recovery by housing associations may be related to the relatively more generous staffing ratios enjoyed by them and to more intensive supervision and monitoring by external bodies. The lower rate of translating decrees for repossession into evictions by local authorities may be because of their more limited staff resources or because they perceive themselves as landlords of last resort, with obligations under the homelessness legislation. The different degrees of involvement of elected and committee members may also be having an impact. However, the extent and nature of the contribution made by each of these factors is hard to judge from the information available to us, and it would appear that further research is required. We are more confident that it is possible to establish a link between tenant profiles and different levels of rent arrears across tenures, but here too, further research is needed to confirm the link.

While our research suggests that the specific differences between legal rights in the two tenures have played a relatively minor role in determining the outcomes of the repossession process this is not an argument for being unconcerned about the legal divide in the social rented sector. The creation of assured tenancies was one amongst a number of measures intended to create a market-dominated housing system. The decision appears to have been taken mainly on ideological and theoretical grounds – the assumption that this was necessary to encourage private sector lenders to support housing association

 © The Modern Law Review Limited 1999

development.[66] Yet this objective was at odds with consumerist moves to extend tenants' rights which meant that assured tenancies did not assist the Government's attempt to transfer housing to the housing association sector *with tenants' approval.*

In practice, a market-dominated system has not been achieved. The housing association sector has retained high levels of public funding, its close relationship to public authorities and its social service ethos. The regulatory regime applying to it has not loosened. In adopting and promoting the MATA, housing associations played a role in creating the conditions which make assured tenancies appear as an anomaly in housing policy. In addition, our analysis of the outcomes of the repossession process suggests that housing associations do not need the greater freedom to evict afforded in assured tenancies in order to ensure that rent arrears are confined within reasonable levels. Research on lenders' perception of risk suggest that lenders are most interested in a steady rental stream, an adequate asset base, and staff expertise.[67] The evidence does not suggest that lenders would see secure tenancies as an impediment to private funding. So, on pragmatic grounds, there is no case for assured tenancies to be applied to housing associations.

The rights of different types of social sector tenants is likely to remain an issue as long as the difference remains, just as the differences in access to social services, health and education within different parts of the public sector have become an issue in the 1990s with the creation of fund-holding practices and opted-out schools, for example. The reasons which justified differences in the social rights of citizenship for sub-groups of the most vulnerable members of society between 1987 and 1997 appear anomalous under a new Labour government. If the idea of local housing companies taking over public sector stock continues to develop[68] the nature of the tenancy rights applying to tenants considering transfer will continue to be an issue. This, therefore, seems a good time to reconsider the question of whether assured tenancies should be used in the social rented sector and whether, as a matter of principle, tenants rights of security of tenure should or need differ.

66 *Housing: the Government's Proposals for Scotland* Cm 242 (1987) para 5.7.

67 R. Chaplin *et al, Rents and Risks: Investing in Housing Associations* (York: Joseph Rowntree Foundation, 1995); D. Robertson, *Housing Associations in Scotland* (Edinburgh: Scottish Homes, 1994). See also HC 1st Scottish Standing Committee col 642 11 February 1988.

68 *Our Future Homes: Opportunity, Choice, Responsibility*, Cm 2901 (1995) and Part I of the Housing Act 1996 (which does not apply to Scotland).

© The Modern Law Review Limited 1999

Company Law and the Myth of Shareholder Ownership

Paddy Ireland*

In recent years, the rather arcane subject of corporate governance, meaning the governance of the public companies that dominate the economy,[1] has risen high on the political and legal agenda. Various reasons for this can be identified, prominent amongst them the debates, with which the governance issue has become entwined, about the virtues in relation to both social welfare and international competitiveness of different versions of capitalism and the corporation. As company lawyers are well aware, diverse opinions have emerged, with some advocating the adoption of a legal model of the company based around so-called stakeholding principles akin to those said to be found in Germany and Japan, while others seek to reinvigorate the traditional, shareholder-oriented, Anglo-American model.[2] Despite these differences, however, there is widespread agreement that shareholders have an important role to play in ensuring good governance. For some, good governance requires a restoration of shareholder supervision and control.[3] For others, including many supporters of 'stakeholding', it should not be judged purely in terms of maximising 'shareholder value' but still requires more 'committed' ownership by shareholders, if only to eradicate the danger of 'short-termism'.[4] In keeping with this, the Labour government has recently asserted the need for more active and less fickle shareholding and has for some time been toying with the idea of making voting at company general meetings compulsory for institutional investors.[5]

Underlying this consensus is a shared assumption: that the shareholders of large corporations 'own' the companies concerned;[6] or in the 'nexus of contracts' or 'agency' theory of the company, in what amounts to the same thing, that the

* Law School, University of Kent.

1 This article uses the terms 'corporation' and 'corporate' to refer to large public companies, and 'corporate governance' to refer to the general control and accountability of corporate executives rather than to issues of day-to-day management. Concern with corporate governance is a largely Anglo-American phenomenon.

2 Compare, for example, John Parkinson, 'Company Law and Stakeholder Governance' and David Willetts, 'The Poverty of Stakeholding' in Gavin Kelly *et al* (eds) *Stakeholder Capitalism* (London: Macmillan, 1997)

3 '[E]ffective, internationally competitive corporate governance requires the efficient discharge of the ownership role', Allen Sykes, 'Proposals for Internationally Competitive Corporate Governance in Britain and America' (1994) 2 *Corporate Governance: An International Review* 187, 194.

4 See, for example, Will Hutton, *The State We're In* (London: Jonathan Cape, 1995).

5 See Margaret Beckett (President of the Board of Trade), speech to PIRC, 4 March 1998

6 Company lawyers, while generally skirting this issue, sometimes acknowledge that shareholders are not 'owners' of the company in the usual sense of the word. They tend to assume, however, that they have a proprietorial interest in the company akin to ownership, hence, for example, the widespread references to 'the separation of ownership and control'. Non-lawyers tend to be less hesitant. In the *Financial Times* (27 April 1998), for example, it was recently suggested that the idea that shareholders own corporations was 'the most basic tenet of the Anglo-Saxon view of capitalism'. Confirming this, a City correspondent commenting recently on the behaviour of the directors of Lonhro, suggested that they had 'forgotten . [the] principle of company law ... that shareholders *own* the firm and directors merely *run* it' (her emphasis) *The Guardian*, 18 April 1998.

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

shareholders own not 'the company' but 'the capital', the company itself having been spirited out of existence.[7] It is a natural corollary of this assumption that the interests of shareholders should take priority, if not complete precedence, over all others; and that shareholders should, as of right, have a substantial, if not an exclusive, say in the running of companies. As a result, stakeholders and others seeking significant governance reform are, in effect, placed in the position of asking shareholders to give up some of their ownership rights, or of trying to persuade them to exercise them in particular ('socially responsible') ways, either by arguing that it will be in their own best long-term interests to do so, or by appealing to their altruism.

However, despite the general acceptance of the 'ownership' assumption, the legal nature of the share and shareholding are surrounded by uncertainty. As L.C.B. Gower says, the share does not readily fit into any 'normal legal category'.[8] Company lawyers are clear what a share is *not* – apart from when a company is wound up, a rare occurrence in the case of a public company, it is not an interest in the corporate assets.[9] But they are much less clear what it actually *is*. The question 'what ... is the exact juridical nature of the share', Gower observes, 'is ... more easily asked than answered'. It is, perhaps, for this reason that 'a definition of shares ... is something which text books have rarely attempted',[10] though given the importance in the modern world of the share as a form of property and its position at the heart of company law, mediating the relationship between companies and their shareholders, this is rather curious. Moreover, as many have pointed out, it is in many ways equally difficult satisfactorily to distinguish shareholders from debenture holders. In 'legal theory' they are 'rigidly separated', but 'in economic reality [they] merge into each other, ... [a] close examination of the rights conferred by [them] show[ing] the impossibility of preserving any hard and fast distinction between them which bears any relation to practical reality'.[11] Rarely are these problems explored in any depth, however, one of the unfortunate effects of the relatively uncritical acceptance of the 'ownership' assumption having been to foreshorten, and substitute for, analysis. This article seeks to lift the carpet of 'ownership' under which they have been swept and, in so doing, hopes to shed light on some of the conceptual and theoretical conundrums that beset modern company law and to clarify some of the issues underlying the governance debate.

---

7 In a company law equivalent of Mrs Thatcher's 'there is no such thing as society ..', supporters of the 'nexus of contracts' theory argue that there is no such thing as the company only the contractual arrangements between the individual actors involved in the firm As I later explain, this circumvents rather than resolves the corporate ownership problem. For an exposition of the contractual theory of the corporation, see F. Easterbrook and D. Fischel, *The Economic Structure of Corporate Law* (Cambridge Mass: Harvard UP, 1991) ch 1, and for an application of these ideas to British company law, see Brian Cheffins, *Company Law: Theory, Structure and Operation* (Oxford: Clarendon Press, 1997). For a critique arguing that 'the company' as constituted by these theories is composed not of a nexus of contracts but a miasmic nexus of metaphors, see David Campbell, 'The Role of Monitoring and Morality in Company Law: A Criticism of the Direction of Present Regulation' (1997) 7 *Australian Journal of Corporate Law* 343

8 Paul Davies, *Gower's Principles of Modern Company Law* (London: Sweet & Maxwell, 6th ed 1997) 299, 321. The most recent edition of Gower was produced by Paul Davies, but the sections referred to in this article are largely unchanged from earlier editions and I will, therefore, attribute the views expressed to Gower rather than to Davies.

9 *Short Bros* v *Treasury Commissioners* [1948] 1 KB 116.

10 (1989) 10 *Company Lawyer* 140

11 Gower, n 8 above, 301. Paradoxically, this 'economic reality', recognition of which ostensibly weakens the claims of the corporate shareholder, finds some expression in 'nexus of contracts' theories of the company which seek to assert the ultimate supremacy of the shareholder interest.

© The Modern Law Review Limited 1999

Usury and the concept of partnership

'Of all the branches of law', Gower writes, '[company law] is perhaps the one least readily understood except in relation to its historical development'.[12] Nowhere is this truer than in relation to the sharp theoretical line drawn within company law between the investors who 'in economic reality merge into each other' – the line between shareholders standing *inside* the company with rights *in* it – and creditors standing *outside* the company with rights *against* it; between those who 'own' and those who are merely 'owed'. For this line can only properly be understood historically, by reference to the impact of the usury laws on the concept, and later the law, of partnership.

Although the term usury eventually came widely to be used to describe any extortionate bargain, in its original sense it involved loaning money for interest. The usurer was thus an antiquated version of what Marx later called the money capitalist, the only significant difference between usurer's capital and interest-bearing or money capital lying not in their form (both entail a movement from M – M^{1}[13]) but in the social relations of production within which the money moves. Frowned upon in Greek and Roman times and completely prohibited in the Middle Ages, the taking of interest on loans was later permitted but rates of return regulated by statute, at which point usury came to be identified with the taking of excessive interest rather than with taking interest *per se*. To the modern mind, with its acceptance of the intrinsic productivity of money, this antipathy to 'investment', to money as capital, seems a rather anachronistic theological prejudice, a quaint, pre-modern, superstition. In this context, the survival of the usury laws into the nineteenth century[14] appears rather odd; an example of law lagging behind economy. However, while the hostility to usury can be traced back to biblical sources[15] and came to constitute one of the principal elements of canon law, it had social and political, as well as religious, underpinnings. These were influentially outlined by Aristotle[16] and remain relevant today.

For Aristotle, exchange was central to the social intercourse which made community, the pre-requisite of human happiness, possible. By providing a standard of commensurability for unlike things, money facilitated exchange and thereby made an important contribution to social bonding. Usury, Aristotle argued, undermined money's ability to perform these functions and subverted the fairness in exchange which was 'the salvation of states'. It was, he claimed, 'unnatural', for money was inherently *un*productive. Lacking genitals – 'organs for generating any other such piece', as Bentham put it[17] – money was sterile and barren; of all the ways of getting wealth, 'money produced out of money ... was the most contrary to nature'. The usurer made money without working for it, taking, in the form of

12 Gower, n 8 above, 18.

13 Marx characterised the underlying motion which typified capitalist production as involving the circular movement of capital from the capitalist through the process of production back to the capitalist. He expressed this circuit as $M - C - P - C^{1} - M^{1}$, where M represents money, C represents commodities (means of production and labour power), and P represents production. The ^{1} represents an expansion in value, so that the difference in magnitude between M and M^{1} marks the capitalist's profit.

14 They were not finally repealed in Britain until 1854.

15 See, in particular, Luke, 6:35 (*mutuum date nihil inde sperantes*, 'lend, hoping for nothing thereby'); in the Old Testament, see Exodus, 22:25; Deuteronomy, 23:19; Ezekial, 18:17; 22:12; Nehemiah, 5:5

16 Aristotle, *Politics* I, 8–10; *Nicomachean Ethics* V, 5.

17 Jeremy Bentham, *Defence of Usury* (1787), in *The Works of Jeremy Bentham, Volume III* (New York: Russell & Russell, 1962) 16.

 © The Modern Law Review Limited 1999

interest, part of the product of the labour of others. Moreover, Aristotle argued, usury was also unnatural in the sense that the acquisition of money tended towards infinity and endless multiplication, encouraging greed for its own sake. One could, like Midas, have an abundance of money and yet still perish from hunger. For Aristotle, therefore, usury undermined justice in exchange, corrupted market relations and subverted the ethical roots of social, political and economic order.[18] There were many later expressions of this view, among the most vivid, that found in *The Divine Comedy*, in which Dante located usurers, their faces charred beyond recognition by fiery rain, in the seventh circle of the *Inferno* at the very edge of the second division of hell, to reflect 'the degree to which [their] particular offense destroy[ed] communal life and the possibility of spiritual happiness...'.[19] In marked contrast to modern legal systems, Dante treated fraud more harshly than violence precisely because it eroded the trust and confidence without which community would disintegrate. The theory of usury which emerged in the Middle Ages was 'not, then, an isolated freak of casuistical ingenuity, but [a] subordinate element in a comprehensive system of social philosophy'.[20]

The legal antipathy to usury can be traced back at least as far as Roman law. However, not only did Roman law permit usury as long as the interest charged was not excessive, it distinguished potentially usurious contracts of loan (*mutuum*) from other transactions in which payment for the use of money was considered legitimate, prominent amongst them the contract of partnership (*societas*).[21] Over time, as trade expanded and money came increasingly to be borrowed as well as lent as capital (rather than to meet an immediate material need), the legitimacy of many forms of 'investment' was established through an expansion of the scope of concepts such as partnership and a corresponding narrowing of the concept of the (potentially usurious) loan. The result was the legal constitution of many inactive providers of money as 'partners' rather than 'lenders', essentially on the grounds that they put their capital at risk. The concept of partnership that emerged – the 'one great and universal form of licit investment in commerce throughout medieval Europe' – entered scholastic thought 'largely in the form given it by Roman law'.[22] Despite their desire to prohibit usury, therefore, for most canonists and jurists 'there was a world of difference between usury, [profit obtained from] a contract of loan ... and justifiable returns derived from partnerships, where there was a

18 See Scott Meikle, 'Aristotle and Exchange Value', in David Keyt and Fred Miller (eds), *A Companion to Aristotle's Politics* (Cambridge Mass: Blackwells, 1991) 156–181.

19 Judith Schenk Koffler, 'Capital in Hell: Dante's Lesson on Usury' (1979) 32 Rutgers LR 609, 619. The usurers were recognisable by virtue of the coats of arms on their purses, suggesting that the usurers that Dante had in mind were not petty consumption lenders but the great international merchant bankers of the day.

20 R.H. Tawney, *Religion and the Rise of Capitalism* (1926; Harmondsworth: Pelican, 1938) 145.

21 See Reinhard Zimmermann, *The Law of Obligations: Roman Foundations of the Civilian Tradition* (Oxford: Clarendon Press, 1996) 153–187, 451–473.

22 John Noonan, *The Scholastic Analysis of Usury* (Cambridge Mass: Harvard UP, 1957) 133–134. There is general agreement that the antipathy to usury influenced the development of the concept of partnership, although the extent of this influence is controversial. In *Studien in der romanisch-kanonistischen Wirtschafts- und Rechtslehre bis gegen Ende des 17. Jahrhunderts* (2 Vols, 1874 and 1883), Wilhelm Endemann argued that although they were essentially loans, certain contractual arrangements (especially *commendas*) were brought by the canonists within their theory of partnership in order to circumvent the usury prohibition, a view later attacked by (among others) W J. Ashley in his *Introduction to English Economic History and Theory* (London: Longmans, Vol 1, 1894) 411–421. Either way, it is clear is that the antipathy to usury, which predated the canonists, exerted considerable influence over the categorisation of money-lending arrangements, including partnerships. See A L. Udovitch, 'At the Origins of the Western Commenda' (1962) XXXVII *Speculum* 198.

© The Modern Law Review Limited 1999

sharing of risk and venture of the capital'.[23] Nevertheless, there was much disagreement (and confusion) among them as to the status of some contractual arrangements in which money was provided in return for a reward, particularly those in which the liability of investors to third parties was restricted.[24] In 'analysing and systematising the law of usury for the first time', however, the canonists not only 'provided a rational foundation for the dramatic growth of commercial and financial life during the Middle Ages',[25] they developed a very spacious theory of 'partnership' based around ideas of risk and ownership. It was encapsulated by Aquinas:

> He who commits his money to a merchant or craftsman by means of some kind of partnership does not transfer the ownership of his money to him but it remains his; so that at his risk the merchant trades, or the craftsman works with it; and therefore he can licitly seek part of the profit thence coming as from his own property [26]

According to this theory, if an investor of money shared the risk of a venture, the arrangement ceased to be a loan and became a partnership. As Patrick Atiyah points out, the concept of risk employed was rather odd for nobody who lends money is ever guaranteed to get it back.[27] Under the theory, however, if, this inherent risk apart, the investor was promised his money back with an additional sum above that legally permitted, the transaction was a loan and potentially usurious. If, on the other hand, the return was neither guaranteed nor fixed in advance but depended on the success, or otherwise, of the venture, the investor was deemed a risk-taking partner. Usury thus entailed the idea of *certain* gain. As Aquinas emphasised, the issue of risk was itself bound up with the issue of ownership. In a loan arrangement, it was argued, the money lender *transferred* ownership of his money to the borrower, took a fixed and guaranteed return, and dispensed with the risk to his property; whereas in a partnership the provider of money *retained* ownership of his property and thus put it at risk. This not only justified his return, it made it possible to attribute it to the goods which the money had been used to buy, rather than to the sterile and barren money itself.[28] In this way, over time, the usury laws contributed to the radical differentiation of two sorts of money investor: 'lenders' *outside* associations receiving interest; and 'partners' *inside* associations sharing profits.

The teaching of the canonists on usury was received in significant part into civil law, and in England, where 'any transaction which involved usury was clearly illegal in medieval common law', as late as the seventeenth century the legal conception of usury remained derivative of canon law.[29] As the economic foundations of society changed, however, there gradually emerged more and more

23 Benjamin Nelson, 'The Usurer and the Merchant Prince: Italian Businessmen and the Ecclesiastical Law of Restitution, 1100–1550' (1947) VII *Journal of Economic History* Supplement, 104.

24 As in the *commenda* (from the Latin *commendare*, meaning to entrust), the precursor of the French limited partnership, the *societé en commandite*. For the canonists, the main problems arose out of pure rentier investment, especially where it was 'riskless' They had no difficulty with arrangements in which investors of money were involved in the management of ventures, for their gains could be attributed to labour and their profits deemed a particular case of wages. See generally Noonan, n 22 above, 133–153; Ashley, n 22 above, 377–488; and John Baldwin, 'The Medieval Merchant Before the Bar of Canon Law' (1959) *Papers of the Michigan Academy of Science, Arts, and Letters* XLIV.

25 Zimmermann, n 21 above, 173

26 Cited Noonan, n 22 above, 143; see also Ashley, n 22 above, 419.

27 P.S. Atiyah, *The Rise and Fall of Freedom of Contract* (Oxford: Clarendon Press, 1979) 130.

28 Noonan, n 22 above, 152.

29 A.W.B Simpson, *A History of the Common Law of Contract* (Oxford: Clarendon Press, 1987) 114–116, 514–516. For details on the reception of canonist principles into English law, see Ashley, n 22 above, 382–385, 460–470

© The Modern Law Review Limited 1999

ways in which the usury laws could be circumvented and more and more methods of investment recognised as legitimate by both the religious and secular authorities, either by official declaration or by acquiescence. With this, the morality and legality of taking a return on money lent was increasingly accepted as long as only a moderate amount of interest and not *turpes usurae* had been charged. In a series of statutes passed in the sixteenth century, English law came effectively to permit the charging of interest of up to 10 per cent per annum;[30] by 1713, however, this had fallen to five per cent, the level at which it remained for the rest of the century.[31]

The result was that during the eighteenth century, the formative period of the modern English law of partnership,[32] the usury laws, still 'feared and ... far from obsolete',[33] continued to influence the characterisation of different forms of 'investment' in English law. Thus, although the law of partnership 'presumed that each partner was an active trader in a joint concern [with] full power to act as agent of his fellow partners',[34] in accordance with the 'risk' theory of partnership the legal status of inactive providers of money vis-a-vis third parties was determined by reference to the content of their financial return rather than by reference to their involvement (or otherwise) in the running of the concern. As Blackstone J explained in *Grace* v *Smith*:

> the true criterion (when money is advanced to a trader) is to consider whether the profit or premium is certain and defined, or casual, indefinite, and depending on the accidents of the trade. In the former case it is a loan (whether usurious or not, is not material to the present question), in the latter a partnership.[35]

Thus 'lenders' who received interest (a return 'certain and defined') were distinguished from 'partners' who received a share in profits (a return 'casual, indefinite and depending on the accidents of trade'), although both received the return on their capital in the same *form* – as a reward for the mere ownership of money. It followed that any investor whose return, even if described as 'interest', varied with the profits of a business was liable to be deemed a partner and fully liable for partnership debts,[36] as were clerks or managers who received a share of profits in return for their services.[37] Partners were also deemed to have an interest in the partnership assets. At common law, they were treated as joint tenants or tenants in common of personal property belonging to the partnership, and as tenants in common of partnership land; in equity, partnership property was treated as though held by all the partners on trust to be used for the proper business

30 For a summary of these provisions, see Simpson, *ibid* 513–518.

31 There were exceptions, for example in relation to the East Indies and the Bank of England. 'Loans' could also be taken outside the ambit of usury by constructing arrangements (such as the 'bottomry bond' and the 'undervalued annuity') in which the principal was 'hazarded', see Sybil Campbell, 'Usury and Annuities of the Eighteenth Century' (1928) 44 LQR 473, 490.

32 The first partnership text, William Watson's *A Treatise on the Law of Partnership*, was published in 1794. Others, by Basil Montagu, Neil Gow and John Collyer followed early in the nineteenth century.

33 See Sybil Campbell, 'The Economic and Social Effect of the Usury Laws in the Eighteenth Century' (1933) *Transactions of the Royal Historical Society* 197, 204.

34 Michael Lobban, 'Corporate Identity and Limited Liability in France and England 1825–67' (1996) 25 *Anglo-American Law Review* 397, 399.

35 (1775) 2 Blackstone 997.

36 *Waugh* v *Carver* (1793) 2 Wm Blackstone 23.

37 *Hesketh* v *Blanchard* 4 East 144. The principal advantage of partnership was that there was no ceiling on the permissible return on investment; the disadvantages the absence of a 'guaranteed' return and, with English law failing to embrace the limited partnership, potential liability to third parties to one's 'last shilling and acre'.

© The Modern Law Review Limited 1999

purposes of the partnership.[38] The law of partnership thus came to embody a distinction similar to that elaborated by Marx between money capitalists, those who contribute funds, and industrial capitalists, those who utilise those funds in the production process.[39] Lenders were, in effect, deemed money capitalists (and, therefore, potential usurers) because they transferred ownership of their property and it did not, therefore, change in form before returning to them: the movement of their capital was $M - M^1$. The capital of partners, on the other hand, even those uninvolved in management, was conceptualised as having moved through a productive circuit, $M - C - P - C^1 - M^1$.[40] Indeed, even in relation to inactive rentier partners, there was some justification for this characterisation. Most early partnerships had been transient affairs, created for a specific trading venture after which assets were sold and proceeds distributed, enabling partners periodically to liquidate their investment and gain access to their capital. By the later eighteenth century, however, as industrial enterprises grew in number, many partnerships were becoming more permanent affairs, with the result that the money of rentier investors was often tied up in productive assets for a prolonged period, giving those investors the character of industrial rather than money capitalists, despite the passive nature of their investment and lack of managerial involvement. Nevertheless, while the usury laws were in force, there was a 'tendency ... to treat what was in truth a loan at usurious interest, and therefore illegal, as a contract of partnership and therefore legal'.[41] Indeed, to some partnership lawyers writing later in the nineteenth century, some of these arrangements involved not true partnerships but loans, suggesting to them that the canonical justification of them and the theory of partnership upon which it was based was deeply hyprocritical.[42]

The personified company and shareholder ownership

In order fully to grasp the significance of these developments for company law, it is important to remember that what we now call 'company law' began life in the early nineteenth century as 'joint stock company law', meaning the body of law applicable to joint stock companies, and that until the latter half of the century, it was considered and treated as an adjunct of the law of partnership. Although founded upon the partnership principle of trading on a joint account, the joint stock company was distinguished from the 'ordinary' partnership on the basis of its economic, rather than its legal, characteristics. Crucially, the rentier investor, who could be accomodated within the ordinary partnership but who was largely peripheral to it, lay at the heart of the joint stock company. Characterised by large numbers of inactive shareholders, by increasingly specialist managers and by

38 This equitable right to have partnership property applied in a proper manner became known as the partner's lien. The modern view is that partners are beneficially entitled to the totality of the partnership assets.

39 'The employer of capital – even when working with his own capital splits into two personalities – the owner of capital and the employer of capital; with reference to the categories of profit which it yields, his capital also splits into capital *property*, capital *outside* the production process, and yielding interest of itself, and capital *in* the production process which yields a profit of enterprise through its function'. Marx, *Capital* Vol III (1894; London: Lawrence & Wishart, 1978) 375

40 Or in the case of merchant's capital, $M - C - M^1$.

41 Nathaniel Lindley, *A Treatise on the Law of Partnership, including its Application to Joint Stock Companies* (London: Maxwell, 4th ed 1878) 22–23.

42 See Parsons, *Principles of Partnership* (Boston, 1889), cited Ashley, n 22 above, 416–417.

 © The Modern Law Review Limited 1999

freely transferable shares[43] – the essential features of the modern 'public company' – joint stock companies could in law be incorporated or *un*incorporated. A joint stock company was a joint stock company irrespective of its legal status.[44] Correspondingly, early 'joint stock company law' texts examined the law relating to both incorporated and unincorporated concerns.[45]

Equally importantly, for many years *all* joint stock companies, incorporated as well as unincorporated, were considered to be types of partnership. In the parlance of the eighteenth and early nineteenth centuries, they were 'public' partnerships, distinguishable from 'ordinary' or 'private' partnerships on quantitative rather than qualitative grounds.[46] Like ordinary partnerships, all joint stock companies tended, therefore, to be conceptualised as aggregates of individuals. Incorporation was seen as creating a separate legal entity, but the resulting 'body corporate' was thought to consist of 'several individuals, united in such a manner that they and their successors constitute but one person in law, a person distinct from that of any of the members, *though made up of them all*...'.[47] There was, as yet, no conception of 'the company', even if incorporated, as something with an existence entirely of its own, as an object *completely* separate from its shareholders. On the contrary, the shareholders *were* the company, hence the persistence for much of the nineteenth century of the view that directors were simply agents of 'the company' and subject to the control of the shareholders ('the company') in general meeting. Hence also the regular description of incorporated and unincorporated companies as 'theys', and the notion, exemplified by the 1856 Joint Stock Companies Act, that people 'formed themselves' into companies, with its implication that companies were made *of*, rather than *by*, them.[48]

Crucially, while they were conceptualised in this way, the principles of the law of partnership, slightly modified, were thought to be applicable to companies, and as the early texts make clear this was considered true of incorporated, as well as of unincorporated, firms.[49] Incorporation was seen as offering joint stocks certain important legal privileges which took them to some extent outside the principles of the law of partnership, but it was not thought to provide a fully fledged alternative legal form.[50] Nowhere was the perception of 'joint stock company law' as a branch of the law of

---

43 See W.R. Scott, *The Constitution and Finance of English and Irish Joint Stock Companies to 1720* (Cambridge: CUP, 1909–12) vol 1 45–46.

44 Charles Wordsworth's *The Law Relating to Railway, Canal, Water, Dock, Gas and Other Companies* (London: Benning, 6th ed 1851) opens with the statement that 'joint stock companies are unincorporate and incorporate'

45 Thus, the first 'company law' text, Charles Wordsworth's *The Law Relating to Railway, Bank, Insurance, Mining and other Joint Stock Companies* (London: Butterworth, 1st ed 1836) covered both incorporated and unincorporated companies. An expanded second edition appeared the following year.

46 See, for example, William Watson, *A Treatise of the Law of Partnership* (London: Butterworth, 2nd ed 1807) 3–5

47 J.W. Smith, *A Compendium of Mercantile Law* (London. Maxwell, 3rd ed 1843) 81.

48 See generally Paddy Ireland, 'Capitalism Without the Capitalist: The Joint Stock Company Share and the Emergence of the Modern Doctrine of Separate Corporate Personality' (1996) 17 *Journal of Legal History* 40.

49 Wordsworth, for example, treats both incorporated and unincorporated joint stock companies as governed by the general principles of the law of partnership except where they had been 'superceded' or 'limited and restrained' by the granting of corporate privileges or status. Wordsworth, n 45 above (2nd ed 1837) 101, 172.

50 'As to the points which belong exclusively to the conception of the business corporation [as opposed to the conception of corporations in general]', wrote Samuel Williston, 'the law has been formed very largely since 1800', 'History of the law of Business Corporations before 1800' (1888) 2 *Harvard Law Review* 105, 113.

© The Modern Law Review Limited 1999

partnership[51] made clearer than in the title of Nathaniel Lindley's *A Treatise on the Law of Partnership, including its application to Joint Stock Companies*, first published in 1860. Indeed, Lindley continued for many years to describe company law as a 'mere statutory development' of the law of partnership[52] and not until the final decades of the century did it finally crystallise into a legal category in its own right. In the meantime, judges regularly turned to the rules of partnership when there was no case on companies to guide them, ceasing 'only when the idea of the legal personality of the company had permeated all the courts'.[53]

In accordance with partnership principles, ownership of a joint stock company share was considered in law to bestow on the holder an interest in the assets of the company, with the shareholders of both incorporated and unincorporated companies conceptualised as the equitable owners of the company's assets, whether realty or personalty. Shareholders were, then, not only 'the company', they were, quite literally, legally constituted as the 'owners' of its assets. In certain respects, this characterisation of shareholders as 'owners' was always rather strained, for as Adam Smith observed, 'the greater part of these proprietors seldom pretend[ed] to understand any thing of the business of the company; ... giv[ing] themselves no trouble about it, ... receiv[ing] contentedly such half yearly or yearly dividend, as the directors think proper to make to them'.[54] Indeed, there remained some ambiguity as to the precise nature and status of joint stock company shareholding, for some companies treated shareholders rather like lenders, paying interest on all paid-up capital from the outset, even in the period before the company had become productive and profitable.[55] However, in other respects there was a material basis for the assertion of a property nexus between shareholders and the company's assets, for although joint stock company shares were formally freely transferable, in theory enabling shareholders to liquidate their capital at will, throughout the eighteenth and early nineteenth centuries the market for shares was primitive.[56] In the absence of a developed share market, the money of shareholders, transformed into assets under the control of the company, was tied up, potentially for an extended period of time,[57] in the same way, though not to the same degree, as the capital of the rentier investor in a partnership.[58] Moreover,

51 Wordsworth refers to joint stock companies as 'public partnerships', regardless of their legal status, and uses the terms 'partner' and 'shareholder' more or less interchangeably. His method of exposition is even more revealing. He presents what he calls 'the general principles of the law of partnership' as they apply to joint stock companies and then seeks to identify deviations from these principles in individual chapters on the companies occupying different fields of activity, see Wordsworth, n 45 above, (2nd ed 1837). Thus, Sealy's suggestion that partnership principles were invoked to fill the 'gaps' in company law is rather misleading. There were, in fact, more gaps than law, with the principles of the law of partnership thought essentially applicable to all joint stock companies. See Len Sealy, 'Perception and Policy in Company Law Reform', in David Feldman and Frank Miesel, *Corporate and Commercial Law: Modern Developments* (London: Lloyds, 1996) 11, 24

52 Lindley, n 41 above, 14

53 Geoffrey Hornsey, 'Alterations to the "Constitution" of Companies' (1949) 61 *Juridical Review* 263, 270.

54 Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776; Oxford: Clarendon Press, 1976 ed, R.H. Campbell and A.S. Skinner) 741.

55 G.H Evans, *British Corporation Finance* (Baltimore: Johns Hopkins, 1936) 76.

56 See Ireland, n 48 above.

57 This was especially true of joint stock companies because of the fields of activity, such as canals and railways, in which they operated. In these areas levels of investment in fixed assets with a lengthy turnover time were high and it was often many years before operations commenced, let alone bore fruit.

58 Thus, 'ordinary shares were viewed as akin to interests in partnerships', Jonathan B. Baskin, 'The Development of Corporate Financial Markets in Britain and the United States, 1600–1914: Overcoming Asymmetric Information' (1988) 62 *Business History Review* 199

 © The Modern Law Review Limited 1999

notwithstanding Smith's disparaging comments, compared to their latter-day counterparts eighteenth and early nineteenth joint stock company shareholders took a much greater supervisory interest in their investments.[59] Some companies even financially penalised proprietors who neglected to attend general meetings in person or by proxy.[60]

The depersonification of the joint stock company and the emergence of autonomous company law

The early-to-mid nineteenth century, however, saw major changes in the economic nature of the joint stock company. The catalyst was a rapid growth in both the number and size of joint stock companies, particularly following the dramatic development of the railway system. Investment in railway companies was not only on a much larger scale than anything previously seen, it embraced groups hitherto uninvolved in investment and took a radically depersonalised rentier form.[61] As a result, in the period after 1830 there emerged for the first time a developed market in joint stock company shares which transformed them into money capital – readily marketable commodities, liquid assets easily converted into money.[62] This transformation in the economic nature of the share was reflected in, and reinforced by, its legal reconceptualisation. Following the seminal 1837 case of *Bligh* v *Brent*,[63] it came to be held that shareholders had no direct interest, legal or equitable, in the property owned by the company, only a right to dividends and a right to assign their shares for value. By 1860, the shares of both incorporated and unincorporated joint stock companies had been established as legal objects in their own right, as forms of property independent of the assets of the company.[64] In constituting the share as a form of property in its own right, in stressing its monied character and in relieving it of any direct connection to the assets of the company, the Courts, in effect, reconstituted the circuit of the shareholders capital as $M - M^1$. Moreover, with this, the capital of joint stock companies seemingly doubled. The assets were now owned by the company and by the company alone, either through a corporation or, in the case of unincorporated companies, through trustees. The intangible share capital of the company, on the other hand, was the sole property of the shareholder. A vital legal space thus emerged between companies – owners of assets – and shareholders – owners of shares.

59 In relation to canal companies, for example, see J.R. Ward, *The Finance of Canal Building in Eighteenth Century England* (London: OUP, 1974). Because the returns on corporate shares were by no means guaranteed, shares tended to be seen more as speculations than investments.

60 See A.B. DuBois, *The English Business Company after the Bubble Act, 1720–1800* (New York: Octagon, 1938) 319.

61 There is a large literature on the financial impact of the railways. See, for example, M.C. Reed, *Investment in Railways in Britain, 1820–1844* (Oxford: OUP, 1975)

62 See Rudolf Hilferding, *Finance Capital* (1909; London: RKP, 1981) 109–110; Paddy Ireland, Ian Grigg-Spall and Dave Kelly, 'The Conceptual Foundations of Modern Company Law' (1987) 14 JLS 149. By this time, the view – propounded by Bentham and the founders of classical political economy, though not by Adam Smith – that the usury laws should be repealed on the grounds that money was a commodity like any other and that trade in it should be free of regulation, was gaining ground. The usury laws survived mounting criticism from these sources in the first decades of the nineteenth century, as well as various parliamentary attempts at repeal, but finally succumbed in 1854

63 (1837) 2 Y & C Ex 268.

64 See Ireland, n 48 above.

© The Modern Law Review Limited 1999

The separation of shareholder from company inherent in the development of the share market[65] and the establishment of the share as an autonomous, liquid form of property was reinforced by other related changes. For example, Sealy suggests that as long as the liability of shareholders remained unlimited 'general meetings were diligently attended and matters of policy were actively debated', and that even after the introduction of general limited liability in 1855 'shareholders continued to take their role seriously'.[66] This exaggerates shareholder assiduity – the great majority of joint stock company shareholders had never in any meaningful sense acted like participating industrial capitalists *inside* the association. Nevertheless, there is no doubt that there was for many years some degree of supervision of company management by shareholders and that this steadily declined as the century progressed. Increasingly, investors moved from holding shares in one or a small number of companies in whose affairs they took an active interest to holding a diversified basket of securities, a development marked from the 1870s by the rise of financial intermediaries and of modern financial reporting.[67] Shareholders sought security not from an active monitoring of management but from the *de facto* limited liability provided by the rise of the fully paid-up share[68] and from risk-spreading. As they thereby became part of what J.H. Clapham called 'blind capital seeking its 5 per cent',[69] any significant involvement in management or managerial supervision more or less ceased. Professional managers were paid to run enterprises, and the great majority of shareholders were reduced to the status of functionless rentiers, receiving their income in the form (if not at the level) of interest – that is, as a return on their capital accruing with the mere passage of time. As the transformation of shareholders 'from active participant[s] to passive investor[s]' was completed, shareholders were not only clearly established (both in law and in economic reality) as money capitalists standing *outside* the company and the production process, the company, the sole legal and equitable owner of the firm's *industrial* capital, was itself 'depersonified ... ceasing to be an association and ... becoming an institution'.[70] Following the 1862 Companies Act, people no longer 'formed themselves' into incorporated companies, they 'formed' incorporated companies, objects external to them, made *by* them but not *of* them. In short, 'the company' was reified.[71]

These changes in the economic (and legal) nature of the company and shareholding had profound effects on the developing law relating to joint stock companies. As the economic and conceptual expulsion of the shareholder was

65 'Shareholders who own equity in a company with publicly traded shares tend not to be tied to the business in the same way as their counterparts in closely held firms ... individuals who buy equity in a public company rarely develop any links to the business by participating in management. As well, when investors decide to terminate the relationship and sell their shares they can usually do so readily', Cheffins, n 7 above, 51

66 Sealy, n 51 above, 25

67 R.A. Bryer, 'The Late Nineteenth-Century Revolution in Financial Reporting' (1993) 18 *Accounting, Organisations and Society* 649; J B. Jefferys, 'Trends in Business Organisation in Great Britain since 1856' (PhD, Univ of London, 1938) ch 9.

68 Until the later nineteenth century, the great majority of joint stock company shares were only partly paid-up. While this remained the case, shareholders remained liable to calls and, in that sense, retained a connection to the company By the turn of the century, however, nearly all shares were fully paid-up. This not only completed the 'externalisation' of the investor, it meant that existing shares ceased to be sources of new capital. See P L Cottrell, *Industrial Finance, 1830–1914* (London: Methuen, 1980) 80–103

69 J.H. Clapham, *An Economic History of Modern Britain* (Cambridge: CUP, 1928) vol 1 388.

70 Sealy, n 51 above, 24–26.

71 See Otto Kahn-Freund, 'Industrial Democracy' (1977) 6 ILJ 65.

© The Modern Law Review Limited 1999

increasingly recognised, so too was the depersonified nature of 'the company' itself. This was reflected both in the Companies Acts 1844–62, which made corporate status with limited liability available on mere registration, and in the common law. The doctrine of ultra vires, for example, which developed principally to protect the integrity of the share as a form of property,[72] placed the newly reified 'company' in certain respects beyond even the unanimous collective will of shareholders.[73] More generally, there was a move from seeing directors as subject to the direction and control of the company, meaning the shareholders in general meeting, to seeing them as a self-standing organ of the company as a separate, depersonified entity. This conceptual change was accompanied by a decline in the right of shareholders to intervene in the day-to-day running of companies and by a steady shift of power from general meeting to board.[74] As their 'ownership' rights were steadily eroded, shareholders 'surrendered a set of definite rights for a set of indefinite expectations'.[75]

In underlining the externality of the shareholder, these economic and legal changes laid the foundations for the emergence of the modern doctrine of separate corporate personality. The '*complete* separation'[76] of company and members that this entailed was not, as company lawyers tend to assume, inherent in the legal act of incorporation.[77] Rather, the legal meaning of incorporation in a business context was reinterpreted in the latter half of the nineteenth century to accommodate the radical *economic* separation of joint stock companies from their shareholders.[78] In heightening the qualitative differences between joint stock companies and 'ordinary' partnerships,[79] these changes also laid the foundations for the emergence of joint stock company law as a body of law in its own right, autonomous from the law of partnership. As Michael Lobban observes, from around mid-century, as 'the company came to be seen as more of an abstract entity

72 See Lord Langdale in *Colman* v *Eastern Counties Railway Co* (1846) 10 Beav 1.

73 *Ashbury Railway Carriage & Iron Co* v *Riche* (1874) LR 7 HL 653.

74 The shift can be seen in the different approaches adopted in *Isle of Wight Railway Co* v *Tahourdin* (1883) 25 ChD 320 and *Automatic Self-Cleansing Filter Syndicate Co Ltd* v *Cuninghame* [1906] 2 Ch 34. Formally, despite the provision of highly influential model articles, British statute law (like that in the US, but unlike that on the continent) does little to mandate how corporations are to be organised, so that corporate governance is, theoretically, largely a matter for the company itself (meaning its shareholders) to determine. This encourages the misleading view, exemplified by 'nexus of contracts' theories, that the corporation is a creature of private contract and that the relation between shareholders and directors is one of principal and agent.

75 A.A Berle and G.C. Means, *The Modern Corporation and Private Property* (1932; New York: Harcourt Brace, revised ed 1967) 244.

76 Gower, n 8 above, 79.

77 Company lawyers tend mistakenly to assume that incorporation carries certain inherent, fixed consequences, among them a 'complete separation' of company and shareholders, arguing that this was not appreciated by early nineteenth century judges. Thus, Gower argues that '...it was not until *Salomon* v *Salomon & Co* at the end of the nineteenth century that [the] implications [of corporate personality] were fully grasped even by the courts', n 8 above, 77.

78 The conceptual problems generated by these processes contributed to the late nineteenth century debates associated with Gierke and Maitland concerning the nature of corporate personality. Mary Stokes has recently revisited these debates, arguing that after 1862 company law vacillated between a contractual and a natural entity theory of 'the company', drawing on the natural entity theory – in which company and shareholders are radically separate – to legitimate limited liability; and on the contractual theory – whereby the company is the product of a contractual agreement between its owners – to endorse the power conferred by shareholders on directors. Eventually, she suggests, faced by the incoherence that resulted from this, 'academics in the early decades of this century adopt[ed] in their writing the natural entity conception of the company'. Mary Stokes, 'Company Law and Legal Theory', in William Twining (ed), *Legal Theory and Common Law* (Oxford Blackwells, 1986) 155.

79 See, for example, *Watson* v *Spratley* (1854) 10 Ex 222; *Baird's Case* (1870) LR 5 Ch App 725 (per James LJ); *Ashworth* v *Munn* (1880) 15 ChD 363.

© The Modern Law Review Limited 1999

and less a collection of equally liable partners', the unsuitability when applied to joint stock companies of partnership principles, with their presumption of active participation and agency,[80] became ever clearer and a 'new company law regime' began to emerge.[81] However, with 'hoary old ideas of partnership continu[ing] to confuse thinking with regard to corporate enterprise', this took time.[82] It was not until the late 1880s, for example, that Lindley finally ceased subsuming his coverage of company law within his treatise on partnership and accorded it a volume of its own.[83]

Company law and the problem of the share

Consideration of the historical circumstances of company law's emergence as an autonomous legal category establishes the close link between the conceptual structure of modern company law and the nineteenth century joint stock company. In the period after 1870, with the rise of the private company, 'company law' came to be applied not only to joint stock companies but to all forms of business organisation, sole traders and small partnerships included, a process legitimated by the famous decision in *Salomon*.[84] As, slowly but surely, it was detached from the joint stock company as an economic form of organisation, 'joint stock company law' became simply 'company law'. However, by this time its basic conceptual structure was already in place, and it was, of course, a conceptual structure designed and developed by legislature and courts with the joint stock company in mind.[85] Despite much subsequent company law reform, this structure has changed remarkably little, generating many conceptual problems. In short, a body of law designed for application to nineteenth century joint stock companies – single entity, national companies whose shareholders had been relieved of any meaningful ownership function – has come to be applied to both small private concerns, in which shareholders and company are often to all intents and purposes one, and to multi-unit, multi-divisional, multi-national corporations. As a result, the conceptual structure of company law has become ever more divorced from the economic realities to which it applies[86] and taken on a life of its own. Detached from its own history and material origins, it tends to be seen as flowing naturally and inevitably from the legal act of incorporation; as existing apart from *any*

80 See Lobban, n 34 above, 399.

81 Michael Lobban, 'Nineteenth Century Frauds in Company Formation: *Derry* v *Peek* in Context' (1996) 112 LQR 287, 288, 312, 318–319

82 B.C. Hunt, *The Development of the Business Corporation in England, 1800–1867* (New York: Russell & Russell, 1936) 129–130. He describes partnership law as 'a barrier to the essential change that had taken place in the position of the .. typical shareholder, [who] was no longer an entrepreneur in the full sense of the word'

83 The fifth edition of his treatise on partnership appeared in 1888, the first edition of his treatise on company law, significantly entitled, *A Treatise on the Law of Companies, considered as a Branch of the Law of Partnership*, in 1889. By that time, a range of well-known company law treatises had appeared, including those of Buckley and Palmer.

84 *Salomon* v *Salomon & Co Ltd* [1897] AC 22 See Paddy Ireland, 'The Rise of the Limited Liability Company' (1984) 12 *International Journal of the Sociology of Law* 239.

85 In the early 1980s, when the government was considering what form a consolidated Companies Act might take, an explanatory note assumed that the formation of a company would naturally be followed by the raising of capital on the market, prompting Sealy to remark that 'the notional company which our lawmakers have in mind [appears] astonishingly, [to be] a company set in the nineteenth century', by which he meant a medium-sized public company. See Sealy, n 51 above, 26.

86 See, for example, Philip I. Blumberg, 'The American Law of Corporate Groups', in Joseph McCahery *et al* (eds), *Corporate Control and Accountability* (Oxford: Clarendon Press, 1993) 305

 © The Modern Law Review Limited 1999

economic reality. This has not only facilitated its manipulation by capital, it has contributed to absurdities – most notably concerning the treatment of parent companies and their subsidiaries[87] – and to conceptual ossification.

Consideration of the historical circumstances of its emergence also reveals company law's continuing failure to come to terms with the material conditions of its own creation. By the end of the 19th century, not only did shareholding no longer entail ownership of the corporate assets, in an era of general limited liability, fully paid-up shares and diversified holdings, it had ceased to be especially risky. Increasingly, shareholders wanted – and corporate managers provided – regular dividends from shares, with the result that shares (and diversified share portfolios) came widely to be seen as sources of steady income streams rather than as speculative financial instruments whose returns varied dramatically with the ups and downs of a business. In short, shares came to exhibit 'debt-like features'.[88] With corporate shareholders ever more 'passive and functionless, remarkable only in [their] capacity to share, without effort or even without appreciable risk, the gains of ... growth...',[89] the justifications for their residual 'ownership' rights became ever harder to discern. And yet, although its emergence as an autonomous legal category was premised precisely on a recognition of the 'complete separation' of shareholder from the joint stock company, company law, 'while stopping short of according shareholders ownership rights over corporations, nevertheless [continued to] vest significant property rights in the shareholders as residual claimants'.[90] It still does, clinging on to the vestiges of shareholder 'ownership' and retaining for shareholders their place at the centre of the governance stage. In this context, one does not have to look far for the sources of the conceptual difficulties that company lawyers have encountered in trying accurately to characterise and define the nature of the share and shareholding.

As Robert Pennington says, 'despite [their] familiarity with them ... [the courts] have found it extraordinarily difficult to define the legal nature of shares'.[91] Indeed, there are marked differences of emphasis even among the minority of company lawyers who bother to attempt to do so. Because of the lack of any direct link between the share and the assets of the company – the term 'share' is 'something of a misnomer, for shareholders no longer share any property in common'[92] – some stress its contractual qualities. Pennington himself, for example, asserts that shares 'are simply bundles of contractual and statutory rights which the shareholder has against the company',[93] approvingly quoting in support the oft-cited definition offered by Farwell J in *Borland's Trustee* v *Steel Bros & Co Ltd:*

> A share is the interest of a shareholder in the company measured by a sum of money, for the purpose of liability in the first place, and of interest in the second, but also consisting of a series of mutual covenants entered into by all the shareholders *inter se* in accordance with [s 14 of the Companies Act 1985]. The contract contained in the articles of association is one of the original incidents of the share. A share is .... an interest measured by a sum of money

87 See P.T. Muchlinski, *Multinational Enterprises and the Law* (Oxford: Blackwells, 1995) 322–345
88 See Baskin, n 58 above, 232–236.
89 J.K. Galbraith, *The New Industrial State* (Harmondsworth: Pelican, 1967).
90 Simon Deakin and Giles Slinger, 'Hostile Takeovers, Corporate Law, and the Theory of the Firm' (1997) 24 JLS 124, 145.
91 Robert Pennington, *Company Law* (London: Butterworths, 6th ed 1990) 56.
92 Gower, n 8 above, 300.
93 n 91 above, 135

© The Modern Law Review Limited 1999

> and made up of various rights contained in the contract, including the right to a sum of money of a more or less amount.[94]

The concept of the share in this decision, Pennington tells us, is clear. The contract contained in the company's memorandum and articles, and in the Companies Act 1985 'gives rise to contractual obligations of each member as regards the company and every other member. The aggregate of these rights and obligations of a member is his shareholding, and when divided between the shares he holds, they constitute his shares'.[95] But Pennington is aware that there are problems with this view, not least because to assert, as he repeatedly does, that shares are simply bundles of contractual and statutory rights held by shareholders '*against* the company' underlines the shareholder's externality to the company and blurs the distinction between them and debenture holders. Thus, he concedes that 'it is tempting to deduce from [the contractual nature of the share] that the relationship between [shareholder] and ... company is that of creditor and debtor'. While, he assures us, this is 'quite wrong',[96] he shows a marked reluctance either to describe the share as 'property' or the shareholder as an 'owner' of the company. The contractual rights which make up the share are of 'a peculiar nature' in that they are transferable and for that reason, he tells us, shares 'have been called "property" '. But he feels unable to endorse this view, merely commenting that it is 'innocuous enough, provided that it is remembered that they do not comprise any proprietary interest in the company's assets'.[97] He describes the conclusion to which he is drawn as 'disappointing', though 'confusing' might be more apt. 'The most that may be said', he says, is that 'shares in a registered company ... are a species of intangible movable property which comprise a collection of rights and obligations relating to an interest in a company of an economic and proprietary character, but not constituting a debt'.[98]

More widespread is the view that shares constitute both property and a proprietorial interest in the company. 'It is tempting to equate shares with rights under a contract', writes Gower, '[but] a share is something far more than a mere contractual right *in personam*'.[99] While it is doubtful 'whether the rights which a share confers on its holder can be classified as "proprietary" in the usual sense', it is clear that 'the share itself is an object of *dominion*, ie of rights *in rem* and not so to regard it would be barren and academic in the extreme'. As he says, for all practical purposes shares are recognised in law, as well as in fact, 'as objects of property which are bought, sold, mortgaged and bequeathed'.[100] More problematic is giving legal substance to the view that the shareholder has a proprietary interest in the company. Although, like Pennington, Gower cites Farwell J's definition of the share, he stresses that while 'it lays considerable and perhaps disproportionate stress on the contractual nature of the shareholder's rights, [it] also emphasises the fact that [the shareholder] has an interest *in* the company'; that 'a share constitutes

---

94 [1901] 1 Ch 279, 288.

95 Robert Pennington, 'Can Shares in Companies be Defined?' (1989) 10 *Company Lawyer* 144.

96 n 91 above, 135–136. A deduction encouraged by company balance sheets which show share capital as an item on the liabilities side together with the debts owed by the company.

97 *ibid* 57.

98 n 95 above, 144. He notes that the definitional problems are not confined to English company lawyers.

99 n 8 above, 300.

100 'They are indeed the typical items of property of the modern commercial era and particularly suited to its demands because of their exceptional liquidity'. This extract is in part taken from Gower n 8 above (4th ed 1979) 400.

 © The Modern Law Review Limited 1999

the holder a member of the company'.[101] He does not, however, find it easy to specify the precise nature and basis of this proprietary interest:

> The theory *seems to be* that the contract constituted by the articles of association defines the nature of the rights, which, however, are not purely personal rights but instead confer *some sort of* proprietary interest in the company though not in its property.[102]

Indeed, Gower expresses concern that an analysis of the 'proprietary and financial aspects of a shareholder's rights' might 'obscure the important fact that his shareholding causes him to become a member of an association'.[103] The problem, as this implicitly suggests, is that the basis for the privileged status of the shareholder as a 'member' is, indeed, 'obscure', hence Deakin and Slinger's observation that 'it is difficult to find support within company law for the widespread assumption ... that shareholders "own" the business...'.[104] Such difficulties multiply when one tries clearly to distinguish shareholders from debenture holders. As Gower acknowledges, the rigid theoretical separation between shareholders, with rights in the company as well as against it, and debenture-holders, with rights against the company but never in the company itself, collapses in contemporary 'economic reality'.[105] Sealy concurs:

> The theoretical differences between being a creditor of the company and being a member are considerable from a legal point of view, but (at least in the case of a solvent and prosperous company) the practical consequences for investors, apart sometimes from tax considerations, are very similar .... an investment in debentures or debenture stock is very similar to an investment in shares: both are securities in the corporate sector of the economy offering different kinds of risk and different kinds of return.[106]

The reality is, of course, that both debenture holders and shareholders are money capitalists, external to companies and to the production process itself. Disinterested and uninvolved in management, and, in any case, largely stripped (in law as well as in economic reality) of genuine corporate ownership rights, the shareholder is, as Berle and Means pointed out, 'not dissimilar in kind from the bondholder or lender of money'.[107] While, therefore, the relationship between shareholder and company is not exactly one of lender to borrower – the share is not, as some have suggested, a kind of loan – neither is it in any meaningful sense one of owner to owned. The share is a particular and distinctive form of money capital; property in the form of a claim on the company's profits.[108]

Corporate governance and the myth of ownership

Ultimately, the conceptual pea-souper enveloping the share and shareholding arises from modern company law's failure to take separate corporate personality

---

101 n 8 above, 301, 299

102 *ibid* 301, emphasis added

103 *ibid* 302.

104 n 91 above, 134.

105 n 8 above, 299–301, 321.

106 L S. Sealy, *Cases and Materials in Company Law* (London: Butterworths, 6th ed 1996) 420–421. This is not, however, to say that there may not be important differences from the corporation's point of view between debt and equity.

107 n 75 above, 245.

108 See David Kelly, 'Money Capital and Company Law' (1990, unpub PhD thesis, Univ of Kent); C.B. Macpherson, 'Capitalism and the Changing Concept of Property' in Eugene Kamenka and R.S. Neale, *Feudalism, Capitalism and Beyond* (London: Edward Arnold, 1975).

© The Modern Law Review Limited 1999

seriously enough, and its consequent attempt to straddle two essentially irreconcilable positions. On the one hand, through the doctrine of separate corporate personality in its modern form, company law expresses and reinforces not only the separate existence of 'the company' but the erosion, *de jure* and *de facto*, of the shareholder's ownership rights. As we have seen, historically these developments provided the basis for company law's crystallisation out from the law of partnership and its emergence as a separate legal category in its own right. The doctrine has also resulted in company law taking separate corporate personality very seriously in certain contexts where it is hard to justify, for example in relation to the treatment of private companies and, more importantly, the subsidiaries of large corporations.

On the other hand, in other ways company law has failed fully to recognise the implications of the depersonification of the company and the reduction of the corporate shareholder to the status of a rentier investor with an interest very similar to that of a debenture holder. It has tried instead to hang on to the (always rather artificial) characterisation of corporate shareholders as 'insiders', 'members' and 'owners', continuing to grant to them exclusive residual 'ownership' rights – most crucially, of course, the right to vote in general meetings. It has done this notwithstanding the true economic nature of the share; notwithstanding the absence of any property nexus between shareholders and the company's assets; notwithstanding the radical externality of shareholders to 'the company' and their superfluousness to and disinterest in the process of production; notwithstanding the fact that there are serious question marks over the legitimacy of their residual control rights, as well as over their desire, competence, and practical ability to exercise them; and notwithstanding the fact that company law itself has done so much to demote them from the status of owners. In short, company law has not taken separate corporate personality *really* seriously in contexts where it would be entirely justified so to do. Fuelled by the ownership myth and the legal remnants which sustain it, it continues rather to treat company and shareholders as in crucial respects synonymous. As a result of this 'anomalous hangover from earlier times', and despite the fact that 'the ... company has ceased to be a 'they' and has come to be seen as an 'it' ', the law insists on treating 'shareholders, collectively, as [its] only legitimate constituency'.[109] Fiduciary duties, in particular, still rest on this 'out-of-date assumption', with English law holding that the duty of directors is to act 'in the best interests of the company', interpreted to mean in the best interests of the shareholders.[110] It is not surprising, then, that when Sealy asks 'how have we lawyers handled th[e] notion that a company is a separate "person"?', his answer is 'very confusingly'.[111] The uncertainty surrounding the nature of the share and

109 Sealy, n 51 above, 26. Elsewhere, Sealy rightly attributes this identification of company with shareholders to the fact that by the time that the complete separation of company and members had been recognised, 'the principles of the director's fiduciary obligations were already well established on the basis that the members collectively *were* the company'. However, he goes on erroneously to suggest that this was, in turn, a product of the influence on the law of unincorporated deed-of-settlement companies in which 'the members *were* the company' The identification of joint stock companies with their shareholders was not confined to unincorporated companies, but was a feature of all joint stocks in the pre-railway age See L.S. Sealy, 'The Director as Trusteee' (1967) 25 *Cambridge Law Journal* 83, 89–90.

110 Modified slightly by s 309 CA 85 which imposes on directors a duty to 'have regard' to employees. The supremacy of the shareholder interest in company law, mirrored in the City Takeover Code, is most clearly revealed, perhaps, by the restrictions placed on company directors (in the name of their duty to shareholders) who wish to frustrate takeover bids. See Deakin and Slinger, n 90 above, 124

111 Sealy, n 51 above, 16.

 © The Modern Law Review Limited 1999

shareholding is just one of the many conundrums that have accompanied the incomplete depersonification of the company.

At present, the tendency among company lawyers is to deal with these problems by discrete omission. The legal nature of the share, if it is addressed at all, is usually afforded cursory treatment, with shareholders simply assumed, in some rather unspecific 'common sense' way, to be the 'owners' of the company.[112] John Parkinson, for example, recognises that 'shareholders are not the owners of the company's assets as a matter of strict law', but quickly adds that 'they are in substance the owners by virtue of being the contributors of the company's capital'.[113] In this context, his argument, although not denying the existence of 'the company', resembles that of 'nexus of contracts' theorists[114] who seek to dispose of the problem of corporate ownership by reducing the company to a series of contractual relations so that it disappears as an entity capable of being 'owned'. The 'ownership of capital', these theorists quite rightly argue, 'should not be confused with ownership of the firm'.[115] However, in disposing of one confusion – that shareholders 'own' the corporation – they substitute another, conflating, under the rubric 'capital', the assets owned by the company and the shares (rights to part of the corporate product, but not to the assets themselves) owned by the shareholders. In dissolving 'the company' and declaring shareholders the owners or 'contributors' or 'providers' of 'the capital', nexus of contracts theorists, in effect, appropriate for shareholders the corporate assets.[116]

Crucially, of course, the 'mistaken analogy'[117] of shareholder ownership, whether of the company itself or of 'the capital', continues to cast a long shadow over the governance debate, serving as the main justification for the anachronistic retention by shareholders of exclusive governance rights and for the claim that public companies should be run predominantly, if not exclusively, in their

---

112 'Ownership' here operates as what Felix Cohen called a 'magic solving word', as a form of 'transcendental nonsense', which curtails discussion of the difficult ethical and practical questions involved See Felix Cohen, 'Transcendental Nonsense and the Functional Approach' (1935) 35 *Columbia Law Review* 809, 820.

113 John Parkinson, *Corporate Power and Responsibility* (Oxford: Clarendon Press, 1993) 34. Thereafter, discussion is relegated to a footnote in which it is noted that 'it can be argued that it is a mistake to regard shareholders as owners at all· they are mere investors ..'.

114 Of whom he is otherwise sharply critical, see John Parkinson, 'The Contractual Theory of the Company and the Protection of Non-Shareholder Interests', in David Feldman and Frank Miesel, n 51 above, 121

115 E F. Fama, 'Agency Problems and the Theory of the Firm' (1980) 88 *Journal of Political Economy* 288, 290

116 Use of the term 'owners' tends to be avoided by these theorists precisely because the company (not the shareholders) owns the corporate assets. However, in dissolving the company as a genuinely separate entity and reducing it to a legal cipher which merely facilitates contracting between different groups, they collapse the distinction between assets and shares, lumping them together as 'capital'. In this way, shareholders, deemed the 'suppliers' or 'providers' of this 'capital', are clandestinely reconnected to the corporate assets, greatly strengthening their bargaining position in the notional contractual negotiations that are meant to take place between the parties involved in the corporation. This characterisation of shareholders as 'suppliers' of capital reveals the underlying nostalgia of these theorists and their reluctance to acknowledge the changes that have taken place in the nature of the corporation and corporate shareholding over the last century and a half. While valid two hundred years ago, the conceptualisation of shareholders as 'suppliers' of capital is now hopelessly outdated, the great majority of shareholders having long been engaged solely in the trading of titles to revenue, not the supply of capital. However, in peddling this sentimental view, which denies the reality of the corporation's autonomous existence and tries to bridge the gap between the corporate assets (owned by the company) and its shares (owned by shareholders), nexus of contracts theory circumvents the awkward questions surrounding the allocation of rights in and over the corporation as an autonomous, property-owning entity. See n 118 and text to n 169 below.

117 John Kay, 'The Stakeholder Corporation' in Kelly *et al*, n 2 above, 125, 131.

© The Modern Law Review Limited 1999

interests.[118] It has also done much to shape the debates about corporate governance, most importantly, perhaps, in causing the issue generally to be cast, following Berle and Means, in terms of a 'separation of ownership and control', with the result that the problem is characterised by many as little more than an administrative dilemma, an agency problem, potentially resolvable by a restoration of shareholder monitoring and supervision. Given the economic nature of corporate shareholding, this solution is not only of questionable legitimacy, it is, as many argue, unlikely to work. Thus, Easterbrook and Fischel deride proposals to enhance shareholder participation[119] and John Kay remarks that 'all experience suggests' that renewed shareholder involvement and supervision is 'not very likely to happen, and would not improve the functioning of corporations if it did'.[120]

As company lawyers are well aware, however, it has forcefully been argued in recent years that the growth in institutional investment and the related rise of the so-called market for corporate control has restored the old model, enabling shareholders effectively to re-assert their rights and, through that, to monitor management and ensure good governance. Despite the divisions of opinion on this matter,[121] it is clear that in the US the evidence of the growing influence of institutional investors is strong,[122] though its forms have changed somewhat in recent years. In the eighties, the emphasis was on take-overs and leveraged buy outs; more recently it has been on what some have called a 'political model of governance'[123] based on the development of longer term investor-corporation relations and generating the rise of what has variously been called a 'fiduciary' or 'investor' capitalism.[124] Although direct and public institutional investor involvement has thus far been less marked in Britain,[125] it is clear that there has

118 Those arguing that the company is a mere 'nexus of contracts' often claim to ground the exclusive corporate governance rights of shareholders not in their 'ownership' of the corporation (the corporation having been eliminated as an entity capable of being owned), but in contract and 'efficiency'. However, like all arguments based on efficiency (and, indeed, contract), these claims presume a pre-existing allocation of property rights, and in relation to the corporation and corporate governance it is precisely this prior allocation that is at issue. Given the transformation (in the wake of the changes in the nature of corporate shareholding) of the corporation from association to institution, how should rights in and over it be allocated? In their anxiety to uphold shareholder privileges, nexus of contracts theorists try to sidestep these issues, effecting a *de facto* allocation of corporate rights by sleight of hand, either by dissolving the corporation, collapsing the distinction between assets and shares, and declaring shareholders owners or 'suppliers' or 'providers' of 'the capital' (see note 132 above); or by arguing that the granting of exclusive governance rights to shareholders is the most efficient arrangement – an argument which is not only contentious (see Gavin Kelly and John Parkinson, 'The Conceptual Foundations of the Company: A Pluralist Approach', forthcoming, *Company, Financial and Insolvency Law Review*), but which uses the concept of efficiency to allocate property rights, rather than to evaluate arrangements based on a prior rights allocation.

119 F. Easterbrook and D Fischel, 'Voting in Corporate Law' (1983) 26 *Journal of Law & Economics* 395, 396.

120 n 117 above, 126.

121 The literature is vast. See, for example, Michael Useem, *Investor Capitalism* (New York: Basic Books, 1996); Margaret Blair, *Ownership and Control* (Washington DC: Brookings Institution, 1995); G. Stapledon, *Institutional Investors and Corporate Governance* (Oxford: Clarendon Press, 1996); Paul Davies, 'Institutional Investors in the United Kingdom' in D D. Prentice and P.R.J. Holland, *Contemporary Issues in Corporate Governance* (Oxford: Clarendon Press, 1993).

122 See, for example, Useem, n 121 above; Doug Henwood, *Wall Street* (London: Verso, 1997) 246–300.

123 J. Pound, 'The Rise of the Political Model of Corporate Governance and Corporate Control' (1993) 68 *New York University Law Review* 1003.

124 On the idea of 'fiduciary capitalism', see R Monks and N. Minow, *Corporate Governance* (Oxford: Blackwell, 1995), and Teresa Ghilarducci *et al*, 'Labour's Paradoxical Interests and the Evolution of Corporate Governance' (1997) 24 JLS 26. On 'investor capitalism', see Useem, n 121 above.

125 See Davies n 121 above; Matthew Gaved, *Ownership and Influence* (London: Institute of Management LSE, 1995).

 © The Modern Law Review Limited 1999

long been 'behind the scenes corporate governance' by financial institutions[126] and evidence is now emerging that UK fund management is becoming more Americanised.[127] Overall, it seems indisputable that the Anglo-American corporate world has changed considerably in the last couple of decades, with corporate conduct becoming 'more focused on rising share prices, and [on making] corporate performance more profitable'.[128] Even, therefore, if attempts by institutional investors directly and in detail to police business policy are the exception and likely to remain so, there can be little doubt that corporate managers are increasingly subject to institutional investor scrutiny and increasingly aware of the importance, not least in terms of their own survival, of maintaining share price and enhancing 'shareholder value'.[129] This is not, of course, to say that increased shareholder activism, with its ultimate sanction of takeover,[130] works so as to weed out 'inefficient' or 'under-performing' managements and to further long-term productive development. On the contrary, there is much evidence suggesting otherwise.[131] As Ghirladucci, Hawley and Williams observe, in the US '[shareholder] activists, when they have been effective, have been concerned overwhelmingly with financial performance rather than long-term productive investment'.[132] The result, paradoxically, is that many now fear that the rise of institutional investment and the imposition on managers of capital market disciplines may be eradicating the 'ownership and control' problem, but in ways that bring a damaging increase in rentier power over corporate policy and threaten a new tyranny of money and finance.[133]

Many are now, therefore, urging that separate corporate personality be taken more seriously. The economist John Kay, for example, favourably contrasts the European and Japanese conceptions of the corporation with those of Britain and America, and advocates recognition of the corporation as a 'public body', as 'an institution with personality, character, and aspirations of its own ... [whose] objectives encompass the interests of a wide-range of stakeholder groups – investors, employees, suppliers, customers and managers ...'.[134] Kay calls for the

126 See John Holland, *The Corporate Governance Role of Financial Institutions* (London: Certified Accountants Educational Trust, 1995).

127 See 'Big investors unite against poor corporate governance' *The Guardian* 23 November 1998. See also *Financial Times* survey, 27 April 1998. Although many of the managers questioned cast doubt on the usefulness and value of the interventions made by fund managers, they confirmed that they had become more demanding The survey suggests that US investors are significantly increasing their UK holdings, which are now conservatively estimated to constitute a 10 per cent stake in the UK market, and that their influence has had 'a far greater impact than the first two governance reports [Cadbury and Greenbury]'. More generally, there is considerable evidence of what Henry Kaufman has called 'the Americanization of global finance', See Henwood, n 122 above, 5. This appears to be reflected in the seeming undermining of the German and Japanese models, see M. Albert and R. Goneno, 'The Future of Rheinish Capitalism' (1996) 1 *Political Quarterly* 184.

128 Charles Craypo, 'The Impact of Changing Corporate Governance Strategies on Communities, Unions and Workers in the U.S.A.' (1997) 24 JLS 10.

129 Meaning total shareholder return: dividends and capital appreciation. The concern with shareholder value has been heightened by the increasingly competitive nature of the institutional investment sector itself.

130 For a brief discussion, see D.D. Prentice, 'Some Aspects of the Corporate Governance Debate', in Prentice and Holland, n 121 above, 36–38.

131 Again, the literature is voluminous. For summaries, see Henwood, n 122 above, 278–282; Deakin and Slinger, n 90 above, 124. The reliance on capital market disciplines is itself a reflection of the decline in many sectors, in the wake of growing oligopolisation and monopolisation, of competitive product markets, see David Campbell, 'Adam Smith, Farrar on Company Law and the Economics of the Corporation' (1990) 19 *Anglo-American Law Review* 185.

132 n 124 above, 29.

133 See Paddy Ireland, 'Stakeholding in the Global Casino' (1997) 24 JLS 276.

134 n 117 above, 126.

© The Modern Law Review Limited 1999

adoption of a German 'stakeholding' conception of the company as 'a community in itself and an organisation in turn embedded in a community', in which directors are cast in the role of trustees of the corporate assets, including its employees. We need, he says, 'an organic model of corporate behaviour which gives to the corporation life independent from its shareholders or stakeholders' and which recognises it 'as an end in itself'.[135] Many others, of course, have similarly advocated the adoption of a stakeholding conception of the company.[136] What is particularly interesting about Kay's argument is that it has led him to reformulate the nature of the governance problem in terms of ownership *per se*.[137] Evaluating the ownership claim of corporate shareholders using A.M. Honoré's analysis of the nature of ownership,[138] he concludes that while they are unquestionably the owners of a corporation's shares, they can hardly be described as the owners of the corporation itself. Of Honoré's eleven ownership tests, they only satisfy two unequivocally and three partially; six are not fulfilled at all.[139] The 'obvious conclusion', he suggests, is that while 'many individuals and groups [customers, shareholders, lenders, employees, directors] have rights and obligations around [public] companies', 'none of these claims can plausibly be described as ownership'.[140] Moreover, Kay is not alone in framing the issue in these terms. The management theorist Charles Handy similarly questions the idea that shareholders are 'owners', likening them instead to racegoers placing their money on financial runners. 'Punters or speculators they may be', he argues, 'owners in any real sense they cannot be'.[141] Although less sympathetic to stakeholding, he, too, argues that we need 'to rethink what we mean by a company' and see it not as property but as a community.[142]

Rethinking corporate shareholding and corporate property rights

The question remains however, how, having recognised the 'existential corporation',[143] does one begin to give content to the idea of the independent corporate interest which underlies it? Like many others, Kay's concern is with furthering the independent *business* interest of the company; with what Lipton and Rosenblum call 'the interest of the corporation in its long-term success as a

135 John Kay and Aubrey Silberston, 'Corporate Governance' (August 1995) *National Institute Economic Review* 84, 86–91.

136 See, for example, Parkinson, n 113 above; Will Hutton, n 4 above. In the US, see the essays in Lawrence E. Mitchell (ed), *Progressive Corporate Law* (Oxford: Westview Press, 1995).

137 He is not alone in doing this. See, for example, Lynne Dallas, 'Working Towards a New Paradigm', in Mitchell (ed), *ibid* 35.

138 A.M. Honoré, 'Ownership' in A.G.Guest (ed), *Oxford Essays in Jurisprudence* (Oxford: Clarendon Press, 1961) 107–147.

139 Honoré outlines eleven characteristics of ownership, the presence of a sufficiency of which would justify the ascription of the term 'ownership': possession, use, management, right to income, right to capital value, right to security, power to transmit, no limit of time on rights, duty to refrain from harmful use, right to use in satisfaction of judgement, right to residual control.

140 Kay, n 117 above, 131 The ownership position of shareholders in private companies is, in reality, if not in legal theory, very different.

141 Charles Handy, *Beyond Certainty* (Boston· Harvard Business School, 1996) 63 Rejection of the perception of shareholders as owners is also to be found among corporate managers, unhappy at growing institutional investor pressure: 'Some [institutional investors]', one disgruntled director recently complained, 'believe that they own the business But they are traders in financial instruments', *Financial Times*, 27 April 1998

142 *ibid* 68–70, Charles Handy, *The Empty Raincoat* (London: Hutchinson, 1994) 143–144, 152–153.

143 Handy, n 141 above, 69.

© The Modern Law Review Limited 1999

business enterprise'.[144] He therefore advocates managers who will further 'the broad purposes of the corporation, and not simply ... the financial interests of shareholders'. For him, the choice between different governance structures 'ultimately depends not on legal theory but on economic performance' and the purpose of the corporate manager should be to 'build a good business' which is internationally competitive.[145] But as Paul Davies points out, this particular attempt to free corporations from shareholder domination risks an unhelpful form of corporate reification. 'The notion of the corporation having an independent interest', he observes, 'serves to obscure the potential conflicts among the various groups of persons affected by the way the company is run, or worse, to disguise a policy of promoting the interests of one of those groups at the expense of the others'.[146] Arguably, the idea of the corporation's independent business interest entails not so much a genuine balancing of the interests of different 'stakeholder' groups, but a tempering of the *immediate* shareholder interest in such a way as to elevate the *long-term* shareholder interest, represented by the independent interest of 'the corporation'. In this context, stakeholding is likely to be endorsed only to the extent that long-term shareholder value is better enhanced by building better relations with employees, suppliers and customers.[147]

Other versions of stakeholding seek a more radical detachment of 'the company' from its shareholders, one which recognises the corporation as a social institution and downplays the idea of it as a private, profit oriented 'business'. Thus, one finds calls for a 'new organisational paradigm' which recognises the 'productive organisation as a public institution with responsibilties which go beyond the residual claims of owners of capital';[148] for a conception of the corporation which embodies an 'enlarge[d] vision of the firm' as a 'political and social institution with its own organisational and industrial relations dynamics'.[149] However, the specific proposals for reform which have thus far emerged are relatively modest. They are generally aimed either at influencing or tempering shareholder behaviour, or at giving greater 'voice' within companies to other 'stakeholder' groups. Thus, many seek to make shareholders more 'committed' and less 'short-termist', through, for example, quinquennial directorial elections which would effectively disenfranchise shareholders in the interim period.[150] Others advocate locking shareholders into their bets by making it harder for them to switch corporate horses, for example by levying taxes on securities trading.[151] Still others propose changes in directors' duties to encompass other corporate stakeholders.[152] More radically, some suggest

144 'A New System of Corporate Governance' (1991) 58 *Chicago Law Review* 187 at 189

145 Kay, n 135 above, 91; n 117 above, 132.

146 Davies, n 121 above, 75.

147 See Paddy Ireland, 'Corporate Governance, Stakeholding and the Company: Towards a Less Degenerate Capitalism?' (1996) 23 JLS 287

148 Hugh Collins, 'Organisational Regulation and the Limits of Contract', in Joseph McCahery (eds), n 86 above, 91, 99. See also the work of John Parkinson, especially n 113 above.

149 See Karl Klare, 'Comment: Untoward Neutral Principles – Market Failure, Implicit Contract, and Economic Adjustment Injuries' (1993) 43 *Toronto Law Journal* 393.

150 Lipton and Rosenblum, n 144 above. For a discussion of this and similar proposals, see Prentice, n 130 above, 28–31.

151 These proposals, associated with James Tobin, are aimed at reducing trading volume and deterring those trying to profit from every twitch in share prices.

152 See, for example, Kay and Silberston, n 135 above, 94. Proposals to reform directors' duties have a long pedigree Sealy argued (many years ago) that a reappraisal of directors' duties was 'long overdue, [for] the directors especially of large companies find themselves in the position of being expected as a matter of business to consider the claims of many competing interests, while being legally answerable to only one', see Sealy, n 109 above, 89–90.

© The Modern Law Review Limited 1999

giving board representation to 'stakeholders', perhaps through the introduction of two-tier boards.[153] However, while many of these proposals would, in different ways, erode still further the dwindling 'ownership' rights of shareholders, they fail fully to confront the ownership issue, leaving relatively untouched what Charles Handy has called the 'extraordinary privilege' accorded to corporate shareholders who, 'for the price of their bets, [are] given a vote from time to time in the auction ring as to who should own their horse', with the result, of course, that they need continuously to be wooed. In a context in which corporate managers, quite apart from their class predilections and share options, are increasingly subject to growing institutional investor, capital market and rentier power, it must be doubted whether attempts to enhance the power and rights of stakeholding groups, alone, will achieve a great deal.[154] Indeed, at present, notwithstanding the rise of stakeholding, we appear to be moving ever further from the realisation of a conception of the corporation as a social institution, for as John Gray points out, with the spread of neo-liberal ideology and contract culture, as an institution, the corporation, like many others, is being 'weakened' and 'hollowed out'.[155]

For this reason, any attempt to realise a fully social conception of the corporation must begin with a fundamental re-evaluation of the nature of corporate shareholding and of shareholder rights. This necessarily entails dispelling not only the myth of shareholder corporate ownership, but other myths that serve to legitimate their governance rights. For example, it is often claimed that shareholders are 'risk-takers' deserving of reward as 'providers' of capital. But this misunderstands both the risks associated with and the nature of corporate shareholding. While equity investment is theoretically riskier than, say, investment in government stock, it is clear that, historically, equities have offered both low-risk investment and, in the medium to long-term, better average returns.[156] Even more importantly, the great majority of share dealings involve not issues raising capital for new investment but the buying and selling of titles to revenue issued long ago. As a whole the stock market today does little to raise capital for new investment. Between 1981 and early 1996, for example, US nonfinancial

153 See Parkinson, n 2 above, 152–153. See also Sally Wheeler, 'Works Councils: Towards Stakeholding' (1997) 24 JLS 44.

154 See Ireland, n 147 above.

155 *False Dawn: The Delusions of Global Capitalism* (London: Granta, 1998) 72.

156 For a summary of the evidence, see Roger Alcaly, 'How to Think About the Stock Market', *New York Review of Books* (June 1998) 22. The risks associated with equity investment have recently been thoroughly aired in the Courts in the context of the formulation of the rules concerning the assessment of damages for future pecuniary loss, an important aspect of which concerns the calculation of the rate of interest which the money awarded will earn. For some years, the practice of the courts was to employ a discount rate of 4–5 per cent to reflect the long-term, after tax, rate of return on equities. This practice was challenged, however, by the Law Commission and the Ogden Report for under-compensating plaintiffs – the rate of return, it was argued, should be calculated on the basis that plaintiffs invest in the least risky securities (index-linked government securities) which would return 2–3 per cent – only for the Court of Appeal to endorse the established view. Damage awards, it held, should be fixed on the assumption that plaintiffs will adopt a 'prudent investment strategy', interpreted to mean that they will invest in a basket of securities including a substantial proportion (about 70 per cent) of equities (see *All ER Annual Review* 1997 538–541 for a summary). The House of Lords has, however, now overturned this decision, holding that plaintiffs are entitled to have their damaged calculated on the basis that they will invest in 'risk-free' ILGS with a low average return of 2.5 per cent. While equities, they argued, were 'the best long-term investment for the ordinary investor', because of possible short-term variations in their value this was not necessarily true for an injured plaintiff needing to live off the capital and income. See *Wells* v *Wells* [1997] 1 All ER 673 (CA); 24 July 1998 (HL). In fact, as we move closer to a world market in capital, and as investors diversify their financial holdings, the levels of return on investments, equities as well as government bonds, seem to be showing clear signs of convergence, see Gray, n 155 above 61–62.

© The Modern Law Review Limited 1999

corporations retired more stock ($700 billion) than they issued, thanks to takeovers and buybacks.[157] It is also important to dispose of the suggestion that the shareholder interest is acquiring 'a new legitimacy' because it now represents, in part, the long-term savings of the more privileged members of the working class.[158] While for some this 'socialisation' of shareholding holds out the prospect of pension-fund socialism,[159] as Doug Henwood points out workers would be ill-advised 'to trade a few extra percentage points return on their pension fund, on which they may draw some decades in the future, for 30 or 40 years of falling wages and rising unemployment insecurity'. Given that the whole purpose of shareholder activism is to increase the profit share of national income and to claim a larger proportion of that profit for rentiers, he explains, 'any gains to people of modest means are accidental'.[160] In short, shareholders should be recognised for what they really are – 'functionless investors'[161], passive owners of claims to part of the labour of others with a resemblance to old-fashioned usurers – and not mistaken for dynamic, risk-taking, deserving, corporate 'owners'.[162]

As Berle and Means noted many years ago, this places a question mark over the legitimate extent of shareholder rights. Having 'surrendered control and responsibility' over corporate assets, shareholders, they suggested, had 'surrendered the right that the corporation should be operated in their sole interest' and 'released the community from the obligation to protect them to the full extent implied in the strict doctrine of property rights'.[163] In fact, what is at issue here is arguably even more fundamental than Berle and Means realised, for the emergence of the modern depersonified corporation represents not merely a dilution of shareholder corporate property rights but the demise of the means of production as private property to which notions of 'ownership', with their connotations of exclusivity and exclusion, are applicable. Indeed, this is something that company law itself recognises, albeit partially, imperfectly and distortedly, in its constitution of the incorporated company as an autonomous, essentially depersonalised, property-owning legal person. It was for this reason that for Marx the rise of the modern corporation represented 'the abolition of capital as private property within the framework of capitalist production itself ... private production without the control of private property'.[164] And that for Schumpeter it '[took] the life out of [private] property', marking the disappearance of 'the figure of the proprietor and with it the specifically proprietary interest' and illustrating how capitalism undermines its own institutional framework.[165] In this context, the share – 'the evaporation of the material substance of property'[166] – is best seen as an attempt to

157 Henwood, n 122 above For this reason, the characterisation of shareholders, which underpins 'nexus of contracts' and 'agency' theories of the company, as 'suppliers' or 'providers' of capital is far too flattering to them and bears little relation to contemporary reality

158 See Davies, n 121 above, 69.

159 See Richard Minns, 'The Social Ownership of Capital' (1996) 219 *New Left Review* 42

160 n 122 above, 293.

161 J.M. Keynes, *The General Theory of Employment, Interest, and Money* (1936), in *Collected Writings* (1936; London: Macmillans, 1973) vol VII, 376. 'Interest', he wrote, 'today rewards no genuine sacrifice'.

162 In this context, perhaps, the need is to distinguish between genuine, socially beneficial, risk-taking which is deserving of reward and gambling which is not.

163 n 75 above, 312.

164 n 39, above, 436. For discussion of the political significance of this, see Shlomo Avineri, *The Social and Political Thought of Karl Marx* (Cambridge: Cambridge UP, 1968) 174–184

165 Joseph Schumpeter, *Capitalism, Socialism and Democracy* (1943; London: George Allen, 5th ed 1976) 139–142

166 *ibid* 142.

© The Modern Law Review Limited 1999

preserve corporate assets as objects for generating private profit even though they are no longer private property 'owned' by shareholders;[167] and the current neo-liberal fashion for theorising the company as a 'nexus of contracts' best seen as an attempt not merely to preserve the share but to legitimate the tenuous claims of vestigial, rentier shareholders to a significant part of the social product. Despite the corporate vanishing trick that these theorists perform, however, the result of their desire to re-establish corporate assets as private property and to forestall recognition of the fact that in the corporate context 'ownership is a problem rather than a pre-supposition'[168] is an ungainly attempt to turn the clock back and to *repersonify* the corporation as an aggregate of capital-supplying individuals.[169]

It is not, however, only the erosion of the means of production as private property and the consequent unsuitability when applied to corporations of traditional ideas of 'ownership' that points to the fundamentally *social* nature of corporate assets. Modern corporations – and the millions who participate in their activities worldwide and the many more millions who depend, directly and indirectly, on the success of those activities – are reflections and expressions of the general decline of production as a 'truly "private" economic activity'.[170] They are aspects of the *socialisation* of production and of the growing economic interdependence that characterises modern capitalism. Increasingly, our collective material fate is inextricably bound up with the use we make of corporate assets. It is arguable, therefore, that not only are corporate assets no longer private property, but that, as the product of the collective labour of many generations upon which we all depend, there are compelling grounds for designating them common property.

Ultimately, it is to these changes in the nature of corporate property and the process of production that stakeholders and others are alluding when they describe the corporation as a 'social' or 'public' institution. In this context, the immediate need is to dispel the residual identification of corporation with shareholders and to complete the process, begun last century, of corporate depersonification. This would not in itself provide answers to the difficult questions concerning the allocation of power and rights in and over corporations and their activities, but it would, by dispelling the ownership myth and the private property premisses upon which it is based, at least enable these questions to be asked.[171] By facilitating recognition of the corporation not as an 'owner', nor as an object capable of being owned, but as a network of social and productive relationships, it would enable us to begin the process of reconceptualising the corporation and corporate property. It cannot be doubted that this – and the replacement of private, shareholder-centred mechanisms by more democratic, social mechanisms of governance that would accompany it – will be extremely difficult to achieve. Quite apart from the formidable political obstacles, as Sealy observes 'at present our company law lacks the conceptual ... tools ... to reflect our new perception of the [public] company as

167 See David Campbell, *The Failure of Marxism* (Aldershot: Dartmouth, 1996) 238–248.

168 Alain Pottage, 'Instituting Property' (1998) 18 OJLS 331, 338

169 It is not, therefore, insignificant that in their attempt to reconstruct corporate relations these theorists rely heavily on the concepts of agency and contract, with the curious result that 'the corporation' as they conjure it up bears a closer resemblance to the personified joint stock company of the eighteenth and early nineteenth centuries (particularly the unincorporated variety) than it does to the modern depersonified corporation.

170 Schumpeter, n 165 above, 141–142.

171 As Joseph Singer points out, it is precisely because we need to resolve difficult issues of this sort that we ask questions about who is owner. See Singer, 'The Reliance Interest in Property' (1988) 40 *Stanford Law Review* 614, 637–641.

© The Modern Law Review Limited 1999

no longer a shareholders' collective, but an enterprise in which the interests of many stakeholders have to be balanced'.[172] We have to start somewhere, however, and we could do worse than to heed Handy's advice and demote shareholders to the status of 'mortgage men'.[173] Not only would this be more befitting their largely parasitical role, it would, by transforming share markets into betting shops rather than auctions and by loosening the grip of capital, help to create more space within which the implicit sociality of the corporation might be realised. In the meantime, with the residual 'ownership' rights of corporate shareholders still in place, corporate property will, through the share, continue to be 'a mutation of private property'[174] and a debased mutation at that. In this context, the fears of the ancient and medieval critics of usury seem ever more prescient, for as rentier investment and finance in its manifold forms asserts ever more forcefully its social power and right to part of the unpaid labour of others, we see growing injustice in exchange, a continuing upward redistribution of income (nationally and internationally), and a disintegration of the bonds that tie society together. In the corporate context, as financial markets demand, in the name of shareholder value and financial prudence, cutbacks in both public and private investment, we see increasing exploitation at work, a disregard for communities and the environment, a shortening of time horizons, and a secular decline in 'real' investment. As Deakin and Slinger rightly say, there is 'no guarantee of compatability between a company law system aimed primarily at the protection of shareholder value and society's interest in maintaining a productive corporate sector'.[175] Or as Keynes put it, 'we cannot as a community provide for future consumption by financial expedients but only by current physical output'.[176] Ultimately, a society guarantees its future only by real physical and social investments. For this reason, the increased shareholder activism sought by many, including the present Labour government, is likely to prove to be the disease that purports to be the cure.

172 Sealy, n 51 above, 28.

173 'Financiers', Handy argues, 'would, in effect, hold mortgages but could only intervene managerially if the business reneged on its payments. Some mortgages would carry no repayment obligation but, instead, a share of the income stream for perpetuity', n 142 above, 152–153.

174 Jeremy Waldron, *The Right to Private Property* (Oxford: Clarendon Press, 1988) 57.

175 n 90 above, 146. Efficiency in the production of profits is distinguishable from efficiency in the production of useful things.

176 Keynes, n 161 above, 104. Elsewhere, Keynes wrote of 'the euthanasia of the rentier, and, consequently, the euthanasia of the cumulative oppressive power of the capitalist to exploit the scarcity-value of capital', *ibid*, 376.

© The Modern Law Review Limited 1999

Universities, Defamation and the Internet

Douglas W. Vick,* Linda Macpherson and Sarah Cooper**

The potential effect of defamation law on discourse over the Internet has attracted considerable comment,[1] in part because a high proportion of the small number of lawsuits arising out of Internet communications have involved defamation claims.[2] If the threat of defamation liability becomes an important means of regulating expression on the Internet, the law of England and Wales[3] is likely to play a central role in this development. London's appeal to libel plaintiffs extends beyond the shores of the United Kingdom, with potential litigants attracted by several substantive advantages afforded claimants by English law[4] and by London's recent history of large damage awards and settlements in libel cases.[5] English courts will usually exercise jurisdiction over cases involving parties who do not reside in Britain so long as the defamatory statement at issue is 'published' (disseminated to others) within England and Wales.[6] Since virtually all statements made on the

* Stirling Media Research Institute, University of Stirling.
** Heriot-Watt University
We would like to acknowledge Gavin Little, Nancy Morris, and Pieternella Roberts for their help in designing the survey used in this study, and Lesley McIntosh for her secretarial assistance

1 See eg C. Counts and C.A Martin, 'Libel in Cyberspace: A Framework for Addressing Liability and Jurisdictional Issues in This New Frontier' (1996) 59 *Albany L Rev* 1083; P. Niehaus, 'Cyberlibel: Workable Liability Standards?' (1996) *U Chicago L Forum* 617; C Waelde and L. Edwards, 'Defamation and the Internet: A Case Study of Anomalies and Difficulties in the Information Age' (1996) 10 *Int'l Review of Law Computers and Technology* 263.

2 Actions have been brought in Australia, the United Kingdom, and the United States. See eg F. Auburn, 'Usenet News and the Law' (1995) 1 *Web Journal of Current Legal Issues*, <http://webjcli.ncl.ac.uk/articles1/auburn1.html> (discussing *Rindos* v *Hardwick*, unreported, Supreme Court of Western Australia, 31 March 1994), C. Dyer, 'Scientist wins out of court damages for Internet libel' *The Guardian* 5 June 1995 (Canadian academic obtained settlement in action commenced in the United Kingdom); J.S. Weber, 'Defining Cyberlibel: A First Amendment Limit for Libel Suits Against Individuals Arising from Computer Bulletin Board Speech' (1995) 46 *Case Western Reserve L Rev* 235, 255 (describing settled claim brought in the United States). See generally C. Gringras, *Laws of the Internet* (London Butterworths, 1997) 91.

3 This article focuses on the law of England and Wales because that legal system attracts the vast majority of defamation actions within the United Kingdom. Leading texts discussing English defamation law include P.F. Carter-Ruck and H.N.A. Starte, *Carter-Ruck on Libel and Slander* (London: Butterworths, 5th ed 1997); C. Duncan, B Neill and R. Rampton, *Defamation* (London: Butterworths, 3rd ed 1995); P Milmo and W.V.H Rogers (eds), *Gatley on Libel and Slander* (London: Sweet & Maxwell, 9th ed 1998)

4 The plaintiff in an English libel case is aided by several rebuttable presumptions and is required to prove relatively little in order to establish a *prima facie* case, while available defences are intricate and often difficult to establish. See generally Carter-Ruck and Starte, n 3 above, 93–160. Increasingly, many of the substantive rules of English defamation law have been criticised for over-deterring expression on matters of public interest. See eg E. Barendt *et al*, *Libel Law and the Media: The Chilling Effect* (Oxford. Oxford UP, 1997) 189–197; G. Robertson and A. Nicol, *Media Law* (London: Longman, 3rd ed 1992) 38–39, E. Barendt, 'Libel and Freedom of Speech in English Law' [1993] PL 449; T. Gibbons, 'Defamation Reconsidered' (1996) 16 OJLS 587, I. Loveland, 'Reforming Libel Law: The Public Law Dimension' (1997) 46 ICLQ 561.

5 These awards and settlements routinely approach or surpass £50,000, and far exceed the amounts recoverable in most other jurisdictions in the world. See D.W Vick and L. Macpherson, 'Anglicizing Defamation Law in the European Union' (1996) 36 *Virginia J Int'l L* 933, 938–968.

6 *ibid* 935, 963 n 182; Gringras, n 2 above, 107–125

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

Internet meet this requirement, jurisdiction over an Internet defamation case nearly always can be founded in the English courts, regardless of the origin of the defamatory comment or the residence of the comment's target.[7]

It is not only the originator of a defamatory statement posted on the Internet who could be sued. Under English law, all persons who communicate a defamatory statement to third parties, regardless of whether they originated the statement or simply repeated or disseminated it, have 'published' the statement and are *prima facie* liable to the plaintiff even in the absence of proof of fault.[8] This liability not only extends to 'primary publishers' – authors, editors, and commercial publishing houses, for example – who exercise direct editorial control over published statements, but also to 'secondary publishers' – including printers, newsagents, booksellers, and libraries – who do not play an active editorial role but do make the offending statement available to a broader segment of the public. The institutions and organisations that provide individual computer users with access to the Internet are, effectively, secondary publishers of defamatory messages communicated over the Internet. These Internet service providers control the computer terminals, or 'servers', that make dissemination of defamatory material possible.[9] For example, the institutions that operate the servers on which World Wide Web sites are stored and which allow third parties to view them play necessary roles in the dissemination of any defamatory messages contained on those sites. Defamatory statements made on e-mail mailing lists or distributed message databases like newsgroups can be communicated to others only with the help of the institutions maintaining the computers on which the statements are posted as well as those through which the statements are transmitted to others. Internet service providers maintain integral links in the distribution chain that makes possible the widespread dissemination of defamatory material for which they are not directly responsible, yet for which they are potentially liable.

This potential liability is vast. According to recent estimates, at least 60 million people in 160 countries currently use the Internet,[10] and a projected 200 million people will use it by 1999.[11] Hundreds of thousands of subscribers post and receive messages on e-mail mailing lists dedicated to specific discussion topics; some 100,000 new messages are posted to newsgroups every day; and the World Wide Web allows individuals or organisations to become 'publishers' who can reach millions of Internet users relatively inexpensively.[12] Most of those who post defamatory messages on the Internet lack the resources to satisfy any damages award obtained in a libel action. The comparatively deep pockets of Internet

7 See Waelde and Edwards, n 1 above, 271–272. This point is illustrated by the case of the Canadian academic who recently obtained a settlement against a Geneva-based scientist in a defamation suit brought in London in connection with statements that were made in a USENET discussion group. See Dyer, n 2 above, 5. For a discussion of the dangers of forum shopping in Internet libel cases, see E.J. McCarthy, 'Networking in Cyberspace: Electronic Defamation and the Potential for International Forum Shopping' (1995) 16 *U Pennsylvania J Int'l Business Law* 527.

8 See eg *Cassidy* v *Daily Mirror Newspapers* [1929] 2 KB 331, 354.

9 See Gringras, n 2 above, 92; N. Braithwaite, 'The Internet and Bulletin Board Defamations' (1995) 145 NLJ 1216 For illuminating discussions of how the Internet works, see J.J. Manger, *The Essential Internet Information Guide* (London: McGraw-Hill, 1995); A. Terrett, 'A Lawyer's Introduction to the Internet' in C. Waelde and L. Edwards (eds), *Law and the Internet: Regulating Cyberspace* (Oxford: Hart, 1997) 13–26. See also *American Civil Liberties Union* v *Reno* 929 F Supp 824, 830–838 (ED Pa 1996) aff'd, 117 S Ct 2329 (1997)

10 See European Commission, *Illegal and Harmful Content on the Internet: Communication to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions* (16 October 1996) 1.

11 *Reno* v *American Civil Liberties Union*, 117 S Ct 2329, 1997 US LEXIS 4037 (26 June 1997) *12–13.

12 See *American Civil Liberties Union* v *Reno*, n 9 above, 834, 835, 837.

© The Modern Law Review Limited 1999

service providers, therefore, may prove attractive to claimants seeking to take advantage of the broad definition of 'publication' in English defamation law.

This article concerns the effects of English defamation law on the policies of one subset of Internet service providers – the universities in the United Kingdom that provide Internet access to their staff and students. Universities are the subject of this study for several reasons. First, they are a substantial yet easily measurable segment of the larger community of Internet service providers. Second, the academic setting is particularly significant when discussing the Internet, since the exchange of scientific and other academic research was the engine that drove the Internet's early expansion. Professors, lecturers, researchers, and university students are, both in terms of numbers of users and frequency of use, among the most important groups posting and accessing information on the Internet (in part because universities have subsidised the costs involved). Third, the academic setting puts the competing values implicated by defamation law into sharp focus. Freedom of speech and the tradition of academic freedom are inextricably linked. The rationales for both overlap: they are supposed to promote the search for truth, respect for the autonomy of individuals, toleration of dissent, and orderly social and political reform.[13] To the extent that the fear of liability for defamation encourages a species of non-governmental censorship, these ends are frustrated. At the same time, academic reputations are particularly vulnerable to injury through the reckless misuse of computer-based communications systems like the Internet.[14] Because the use of such communication systems has become so commonplace within academia, reckless messages will reach precisely those whose opinions matter most to the academic defamed: his or her peers.

This article first describes the potential liability of universities for the harms caused by the transmission of defamatory matter in cyberspace. It then examines the policy responses universities have made to the pressures placed upon them to regulate the use of their computer services, focusing in particular on how defamation law has affected those policy responses. The data reported and analysed are derived from a survey of the computer service departments of United Kingdom universities conducted in the summer of 1997. The article concludes with some observations on a time-honoured conflict, arising in a new context: the conflict between academic freedom and the spectre of institutional censorship encouraged by legal rules that make private institutions legally responsible for the communicative acts of others.

The potential liability of universities

A university could be held liable for the dissemination of defamatory material either as a primary publisher or as a secondary publisher. Responsibility as a primary publisher depends on the operation of the vicarious liability doctrine, while responsibility as a secondary publisher largely depends on the operation of the innocent dissemination defence recently codified as section 1 of the Defamation Act 1996.

13 The leading rationales for protecting freedom of speech are canvassed in F. Schauer, *Free Speech. A Philosophical Enquiry* (Cambridge. Cambridge UP, 1982), W.J Waluchow (ed), *Free Expression Essays in Law and Philosophy* (Oxford: Clarendon Press, 1994); K Greenawalt, 'Free Speech Justifications' (1989) 89 *Columbia L Rev* 119

14 See N. Braithwaite and R Carolina, 'Multimedia Defamation' (1994) 12 *Int'l Media Law* 19, 19.

© The Modern Law Review Limited 1999

Liability as a primary publisher

English law places primary responsibility for the harms caused by defamatory statements on the authors and editors of those statements. The individual who initially posts a defamatory message on the Internet would be considered the statement's 'author' and held answerable for the damage it causes. In addition, a moderator of a mailing list or distributed message database who reviews messages before allowing them to be distributed to others could be considered an 'editor' responsible for any libellous allegations those messages contain.[15] If the originator or moderator is an employee of a university, the well-established principle that employers are vicariously liable for the torts committed by their employees in the course of employment could make the university a primary publisher of the defamatory statements.[16] In other words, the university could be liable regardless of whether it exercised control over the publication of such statements, provided the employee is acting within the scope of his or her employment when posting messages or moderating discussion groups.

It is notoriously difficult to determine with precision the circumstances in which tortious behaviour will be regarded as falling within the scope of employment. The courts have hesitated to prescribe any conclusive test, preferring to examine each case on its own facts. One court, however, has suggested that the proper inquiry is whether the employee's conduct is 'part and parcel of the employment' (even if it is 'unauthorised and prohibited') or whether it is 'so divergent from the employment as to be plainly alien to and wholly distinguishable from the employment'.[17] Applying this standard, it is easy to imagine circumstances in which Internet communications by university employees could be regarded as within the scope of employment. For instance, if a university lecturer posts a defamatory message on a mailing list or newsgroup dedicated to the lecturer's area of academic interest, the posting may be considered job-related and the lecturer's university could be liable. Similarly, monitoring discussion groups, sending e-mails to colleagues, and posting information to Web pages could be considered aspects of an academic's diverse and often amorphous role of teacher and researcher, and certainly not 'alien or distinguishable from' the academic's employment. Significantly, a simple policy statement prohibiting defamatory communications on the Internet would be ineffective in alleviating the university's risk of being held vicariously liable as a primary publisher.

---

15 *ibid* 20

16 See W. Bowstead and F.M.B. Reynolds, *Bowstead and Reynolds on Agency* (London: Sweet & Maxwell, 16th ed 1996) 506, 520. See also *Austin* v *Mirror Newspapers Ltd* [1986] AC 299, 316 (Privy Council); *Citizens Life Assurance Co Ltd* v *Brown* [1904] AC 423. Some Scottish cases have interpreted 'scope of employment' more narrowly in defamation cases than in cases involving other kinds of tortious conduct. See eg *Mandelston* v *North British Railway Co* 1917 SC 442, 445 ('The general rule is that the person who utters the slander is alone responsible for it; and it is only in very exceptional circumstances that a principal is liable for the language used by his agent'). This tendency is not readily apparent in English cases, however. See eg *Pamplin* v *The Law Society* (QBD, 13 Feb 1985) (unreported), available in LEXIS, ENGGEN library, CASES file ('the agent or servant will be deemed to have acted within the scope of his employment when the publication or some part of it, although itself unauthorised, is so directly incidental to some act he was authorised to do that it may be said to be a mode, although an improper mode, of doing an authorised act'). Compare *Colonial Mutual Life Assurance Society Ltd* v *Producers' and Citizens Co-operative Assurance Co Ltd of Australia* (1931) 46 CLR 41 (Aus) (insurance company held liable for defamatory statement about another insurance company made by a canvasser employed to solicit business, even though the canvasser's contract forbade him from defaming another person or institution); cf *Aspro Travel Ltd* v *Owners Abroad Group plc* [1995] 4 All ER 728 (assuming without discussion that company would be responsible for employee's statements under similar facts).

17 *Harrison* v *Michelin Tyre Co Ltd* [1985] 1 All ER 918, 920 *per* Comyn J.

© The Modern Law Review Limited 1999

Liability as a secondary publisher

A university's risk of being found vicariously liable for defamatory statements pales in significance when compared to its potential liability as a secondary publisher of defamatory statements communicated through its servers. The magnitude and precise contours of that danger remain unknown. In part, this is because it is unclear that defamation lawsuits arising out of Internet communications will become commonplace. Thus far, Internet libel cases have been rare,[18] and given the expense of libel litigation they may remain so, making Internet libel a relatively unimportant consideration guiding university policy decisions regarding the Internet. It has been suggested, however, that the dearth of Internet libel actions is but an 'artificial hiatus' that will not 'last forever, especially as commercial enterprise on the Web becomes more prevalent'.[19] But the uncertainty universities face is only partly due to the murky prospects of a litigation explosion. Equally important is the lack of clarity in the law governing the liability of secondary publishers, and particularly the equivocal language of the innocent dissemination defence provided by the Defamation Act 1996.

The rules that allow plaintiffs to recover against secondary publishers even in the absence of proof of malice, recklessness, or negligence have long been regarded as overly harsh. In theory, secondary publishers would need to inspect the contents of all materials that passed through their hands in order to identify and remove those that are libellous, an obligation which, when imposed on libraries, bookstores and newsstands in particular, was deemed unreasonably burdensome. The common law's response to the perceived unfairness of holding secondary publishers liable under such circumstances was the innocent dissemination defence,[20] which was recently codified as section 1 of the Defamation Act 1996.[21] While the defence cannot be asserted by primary publishers,[22] it is available to a non-exhaustive list of secondary publishers.[23] This list includes those who are 'only involved' in 'operating or providing any ... system or service by means of which the [defamatory] statement is ... made available in electronic form' (section 1(3)(c)),[24] and those who are 'only involved' as operators or providers 'of access to a communications system', provided they had 'no effective control' over the person primarily responsible for posting or transmitting a defamatory statement over the

18 See Waelde and Edwards, n 1 above, 283–284.

19 L. Edwards, 'Defamation and the Internet: Name Calling in Cyberspace' in C. Waelde and L. Edwards (eds), *Law and the Internet: Regulating Cyberspace* (Oxford: Hart, 1997) 195.

20 *Emmens* v *Pottle* [1885] 16 QBD 354, 357; *Viretelly* v *Mudie's Select Library Ltd* [1900] 2 QB 170, 179.

21 This statutory defence is discussed in L. Dolding and S. Dzioban, 'Electronic Communication and the Defamation Act 1996: Clarity or Confusion?' (1997) 6 *Information and Communications Technology Law* 55; Edwards, n 19 above, 193–195.

22 See Defamation Act 1996, s 1(1)(a). See also *ibid* s 1(2).

23 *ibid* s 1(3). This list includes printers, the manufacturers, distributors, and exhibitors of films and sound recordings, and broadcasters of live radio or television programmes who broadcast defamatory statements 'in circumstances in which [they have] no effective control over the maker[s] of the statement[s]', *ibid* s 1(3) (a)–(b), (d). The list set out in section 1(3) is offered for guidance only; anyone who is *not* the 'author', 'editor', or '[commercial] publisher' of a defamatory statement (as those terms are defined in section 1(2)) can claim the new statutory defence, *ibid* s 1(1)(a)

24 As originally proposed by the government, this subsection was intended to protect those who process or operate electronic communications equipment, such as CD Roms. The language quoted in the text was added to the Defamation Bill while it was being considered in the House of Commons to ensure that providers of electronically recorded information, such as Reuters and other press agencies, would be covered by the Act. See HC Standing Committee A col 4 6 June 1996.

© The Modern Law Review Limited 1999

system (section 1(3)(e)).[25] Either of these subsections arguably offers some protection to Internet service providers, and there is evidence in the legislative record that the government intended service providers to fall under section 1(3)(e).[26]

Like its common law predecessor, the statutory innocent dissemination defence is not absolute. To avoid liability, the defendant must prove that 'he did not know, and had no reason to believe, that what he did contributed to the publication of a defamatory statement',[27] and that 'he took reasonable care in relation to [the statement's] publication'.[28] The Act lists three factors that should be considered in determining whether the defendant acted reasonably: the degree of the defendant's responsibility 'for the content of the statement or the decision to publish it', the circumstances surrounding the publication, and 'the previous conduct or character of the author, editor or publisher' primarily responsible for publishing the statement.[29]

A fundamental problem raised by the language of the new defence concerns the degree of control universities can or should exercise over the information posted on and accessed from the Internet using university computer systems. Increasingly, lawmakers and parents expect service providers to 'regulate' access to offensive material available on the Internet and to control the conduct of those who use their systems.[30] While the sheer volume of information posted on the Internet makes it unfeasible to prevent transmission of all offensive material through university servers, it is possible that universities could exercise a modicum of control over the communicative activities taking place with the aid of their computer systems. For instance, universities could delete defamatory messages posted on Web sites maintained on their servers if they become aware of them.[31] They could use recently-developed software designed to block access to Web sites and distributed message databases containing obscene, indecent, abusive, or racist content,[32] or block access to Internet sites that have been assigned certain 'ratings' by some 'neutral' arbiter indicating that the sites contain an unacceptable level of offensive material.[33] Additionally, if universities are made aware that certain sites or databases frequently contain offensive messages, perhaps through complaints they receive, they could act to block access to those sites or databases. Universities

25 Among others, this provision apparently covers telephone and mobile phone companies and the operators of radio communication systems.

26 HL Deb vol 570 col 605 8 March 1996.

27 Defamation Act 1996, s 1(1)(c).

28 *ibid* s 1(1)(b).

29 *ibid* s 1(5). These factors were influenced by limitations placed on the innocent dissemination defence at common law.

30 For example, the European Commission has encouraged the co-operation of Internet service providers in developing systems of self-regulation. See European Commission, n 10 above, 8–10. The United States Congress has also invited Internet service providers to voluntarily restrict access to 'obscene, lewd, lascivious, filthy, excessively violent, harassing or otherwise objectionable' material by exempting them from legal responsibility for the content of offensive material that remains accessible despite their good faith efforts to block or delete it. See 47 USC s 230(c) (enacted by the Telecommunications Act of 1996, Pub Law No 104–104, 110 Stat 56, 138).

31 See Gringras, n 2 above, 105. The same is not true of messages sent by e-mail, which are difficult to withdraw once they have been transmitted. See Waelde and Edwards, n 1 above, 267.

32 See J. Delacourt, 'The International Impact of Internet Regulation' (1997) *Harvard Int'l LJ* 207, 229–234 Examples of such software include Cyber Patrol, CYBERsitter, Net Nanny, and SurfWatch.

33 See generally C.D. Martin and J M. Reagle, 'An Alternative to Government Regulation and Censorship: Content Advisory Systems for the Internet' (1997) 15 *Cardozo Arts and Entertainment LJ* 409. The Platform for Internet Content Selection (PICS) is a recent attempt by commercial Internet service providers to establish such a rating system. See *ibid* 414–417.

© The Modern Law Review Limited 1999

could also instruct their computer services departments to conduct random searches of the Internet and block access to 'objectionable' Web sites and newsgroups. The usefulness of some of these strategies in preventing the dissemination of defamatory messages can be questioned; however effective screening software and rating systems may be in identifying sites dedicated to pornography or extremist political groups, libellous messages are often contextual and posted in 'reputable' sites, limiting the efficacy of existing blocking technology in the defamation context.[34] Nonetheless, it is conceivable that access to sites or databases containing abusive words could be blocked, as could access to newsgroups or bulletin boards on which users persistently engage in vitriolic exchanges.

With regard to the conduct of the users of university computer systems, several options present themselves. Universities could provide staff and students with training programmes emphasising the potential legal problems posed by the United Kingdom's defamation laws. Journalists are expected to learn the basic principles of defamation law; perhaps users of the Internet, who sometimes reach audiences exceeding the circulation of many newspapers and magazines, should also grasp these basic principles, and should be provided with instruction concerning the applicability and risks of defamation law.[35] Universities could also adopt codes of conduct for users, compliance with which could be voluntary or coerced through the threat of disciplinary sanctions. A more intrusive option might involve the random monitoring of e-mail messages, newsgroup messages, and Web publications posted by students and staff using university computers, or the non-random inspection of messages posted by individuals who have been the subject of complaints or otherwise raised the suspicions of computer services personnel.

The conundrum confronting university computer service departments is that the law provides ambiguous and possibly inconsistent guidance regarding which of these options, if any, they should exercise. The innocent dissemination defence, for instance, is unavailable to those that exercise 'editorial or equivalent responsibility for the content' of defamatory material they disseminate,[36] but is available to those who are '*only* involved' as the operator or provider of communication systems.[37] If a university restricts access to obscene or abusive postings on the Internet, can it still claim to be 'only involved' as a provider of a communications system, or is it exercising a form of editorial control? If the latter is true, universities must choose between exercising virtually no oversight of their users or the material transmitted through their servers, or sacrifice their status as secondary publishers by succumbing to the pressures placed on them to control undesirable content available over the systems they maintain. The 'all or nothing' approach suggested by this dichotomy is not without precedent. In *Cubby Inc* v *CompuServe Inc*, a United States district court found that a service provider that exercised no control whatsoever over the material accessed by its subscribers should be classified as a secondary publisher.[38] In *Stratton Oakmont Inc* v *Prodigy Services Co*, however, another US court found that because the service provider took steps to screen postings for offensive language and to enforce guidelines issued to its subscribers

34 See Wacklde and Edwards, n 1 above, 280.
35 See *ibid* 284.
36 See Defamation Act 1996, s 1(2) (defining 'editor').
37 *ibid* s 1(3)
38 776 F Supp 135, 140 (SDNY 1991) The law of the United States draws a roughly similar distinction between primary and secondary publishers as does the law of England and Wales. See Restatement (Second) of Torts s 581(1) and comment d

 © The Modern Law Review Limited 1999

by deleting offensive or insulting messages, it was exercising a degree of editorial control sufficient to make it a primary publisher.[39] Although *Cubby* and *Stratton Oakmont* attracted considerable criticism and no longer reflect the law of the United States,[40] they were the leading cases addressing the liability of Internet service providers when Parliament was debating the Defamation Act 1996.

These cases suggest that if universities exercise a measure of control over the material transmitted through their servers or the conduct of their users, they risk being classified as primary publishers. However, if universities turn a blind eye to the uses made of their systems, they may run afoul of another basic prerequisite of the innocent dissemination defence: that they take 'reasonable care in relation' to what is published on their systems.[41] At common law, the innocent dissemination defence was developed because it was perceived to be unreasonable to expect libraries, newsagents, and booksellers to screen the contents of every publication they distributed. It seems unlikely that Internet service providers will be expected to do what was not required of distributors of printed publications. However, if there are options available that could reduce the amount of abusive material appearing on the Internet, or at least restrict access to such material, the unanswered question is whether it is 'reasonable' to refrain from exercising some or all of these options. Does the reasonableness requirement mandate training programmes, or the use of screening software, or the blocking of objectionable sites, or the monitoring of computer users? Would inaction in the face of complaints about the content of particular Websites or distributed message databases, or against individual users of university computer services, be deemed unreasonable?

A related set of questions is raised by language in the Defamation Act indicating that the innocent dissemination defence may not be available to service providers unless they have 'no effective control' over the originator of a defamatory statement.[42] The Act provides no guidance for determining when computer users are beyond the 'effective control' of service providers. It may be that a service provider's responsibility for the actions of those within its 'effective control' is coextensive with its responsibility for the actions of agents and employees under the vicarious liability doctrine. However, it may also be that this liability extends beyond the limits of that doctrine. This question is of particular importance to universities, who would not be vicariously liable for libels promulgated by students or staff acting outside the scope of their employment, but who theoretically could exercise some control over these users of their systems. Does the very fact that a university *could* institute a training programme, screen messages, and discipline abusers of computer privileges mean that those who use the university's computer systems are within the university's 'effective control'?

---

39 23 Media L Rep (BNA) 1794 (NY Sup Ct 25 May 1995) This case did not involve a statement made on the Internet, but rather statements made on an information service available exclusively to the subscribers of Prodigy, a commercial Internet service provider.

40 See 47 USC s 230(c) The legislative history of this provision indicates that Congress wished to overrule *Stratton-Oakmont* by allowing Internet service providers to retain their status as secondary publishers even if they make good faith efforts to restrict access to objectionable material M. Meyerson, 'Ideas of the Marketplace: A Guide to the 1996 Telecommunications Act' (1997) 49 *Federal Communications LJ* 251, 285–287. See also *Zeran* v *America Online Inc*, 129 F2d 327 (USCA 4th Cir, 1997).

41 Defamation Act 1996, s 1(1)(b) and (c).

42 *ibid* s 1(3)(e) This subsection was originally intended to apply to Internet service providers See text at n 26 above Arguably, service providers also fall within the definition of subsection (3)(c), which does not expressly require that the users of that system or service are beyond the 'effective control' of the defendant.

© The Modern Law Review Limited 1999

In sum, the terms of the innocent dissemination defence seem to ask universities to walk a tightrope, exercising a level of intervention sufficient to show that 'reasonable care' has been taken, but not so much as to take them outside of the defence altogether. As Lillian Edwards observed, this is the defence's 'inherent catch 22'.[43] If it is accepted, however, that the defence is intended to encourage 'best practice' among service providers,[44] it is likely that if and when called upon to apply the defence to Internet service providers, courts will look to the practice of service providers generally in judging the conduct of a particular service provider. The behavioural response of service providers to the problem of defamation on the Internet, therefore, will not only be influenced by the ambiguities of the law, but also will itself influence judicial interpretation of the law. The next section of this article describes an empirical study of that behavioural response by the university sector.

The universities' response

During the summer of 1997, a questionnaire was sent by e-mail or post to the directors of the central computing or information service departments of all universities in the United Kingdom. Through the inclusion of both open- and closed-response eliciting questions, the questionnaire sought information about the existence and content of codes of practice relating to the use of university computer systems; the training and instruction universities provide staff and students who use university computer systems; the number of complaints about allegedly defamatory statements universities have received, and the policies in place for handling those complaints; policies regarding the denial of computer services for the misuse of university computer systems; whether access to any of the material available on the Internet was restricted; and whether universities as a matter of policy monitored the communications of their staff or students.

The results of the survey are set out below. In analysing the data obtained, we have taken several variables into account. We have looked at whether questionnaire responses differed depending on the status of the responding institution. A university's 'status' is difficult to measure in any scientific way; it is largely a subjective notion, complicated by the variations in quality and reputation of individual departments within United Kingdom universities. Nonetheless, a very rough approximation of the relative status of a university has been made by looking at the date on which an institution achieved university status, since institutional 'maturity' and reputation roughly correlate with one another. On this basis, universities have been organised into three groups. Group A consists of those institutions granted university status before 1963, when the Robbins Report[45] was published; Group B contains the institutions granted university status after the Robbins Report but before the passage of the Further and Higher Education Acts 1992; and Group C is made up of the former polytechnics and colleges that achieved university status after the 1992 Act. Other variables that we have considered include the number of individuals using a university's computer systems, and the ratio between the number of staff and students who use university computer services and the size of the university's computer services support staff.

43 Edwards, n 19 above, 194.

44 See L Davies and C. Reed, 'The Trouble with Bits – First Steps in Internet Law' (1996) JBL 416, 429.

45 *Report of the Committee on Higher Education* Cmnd 2154 (1963)

© The Modern Law Review Limited 1999

In the institutions with the highest ratios, each support staff member had nominal responsibility for more than 1,000 users, while in institutions at the other end of the scale there was a support team member for every 23 users.

In total, 93 computer service departments were sent questionnaires, and 43 responses were received, a response rate of 46 per cent. The response rate from the universities in Group B was in line with the sample as a whole (46 per cent), while Group A was more highly represented (56 per cent) and Group C under-represented (40 per cent). User-to-support staff ratios were generally higher in Group C universities, and it is possible that support staff in these institutions were under greater work pressure and therefore less able to spend the time necessary to complete the questionnaire.

Apart from random non-response – the unavoidable fact that many people are unwilling to take the time to fill out questionnaires, regardless of their subject matter – any systematic differences between those who returned a completed questionnaire and those who did not could be a source of bias. For example, one obvious source of bias is that those actively interested in the issues addressed in the questionnaire are more likely to participate, making the results obtained potentially unrepresentative of the computer use policies followed by United Kingdom universities on the whole. Because the questionnaire was sent to computer service departments, the staffs of these departments might be more likely to have an active interest in the issues where they also play an active part in the policy making process; indeed, around two thirds of the responses came from universities where computer services played a consultative or decision-making role in determining computer use policy. In addition, universities that have already experienced complaints about defamatory material published on their networks might have a more acute interest in the issue than those that have not had to confront this problem, thus skewing the results toward a conclusion that defamation is a more common problem for United Kingdom universities than is actually the case. On the other hand, it is also possible that an institution's non-response might have been influenced by a desire not to prejudice pending or threatened litigation arising out of statements or images transmitted through their computer servers. Finally, although those surveyed were assured that all responses would be kept confidential, some non-respondents may have decided not to participate because they feared that highly sensitive institutional practices such as monitoring might be revealed.

Besides these systematic explanations for non-response, the timing of the survey must be considered. Universities still have relatively little experience coping with the problems raised by Internet communications, and section 1 of the Defamation Act 1996 had been in place for only about a year at the time the survey was conducted. One respondent indicated that policies concerning many of the issues raised in the questionnaire were still being developed, and it may be that institutions without formal policies were less likely to respond, another potential source of bias in the study. The results of this survey, then, are but a (perhaps somewhat out-of-focus) snapshot of a sector that is still in a transitional stage.

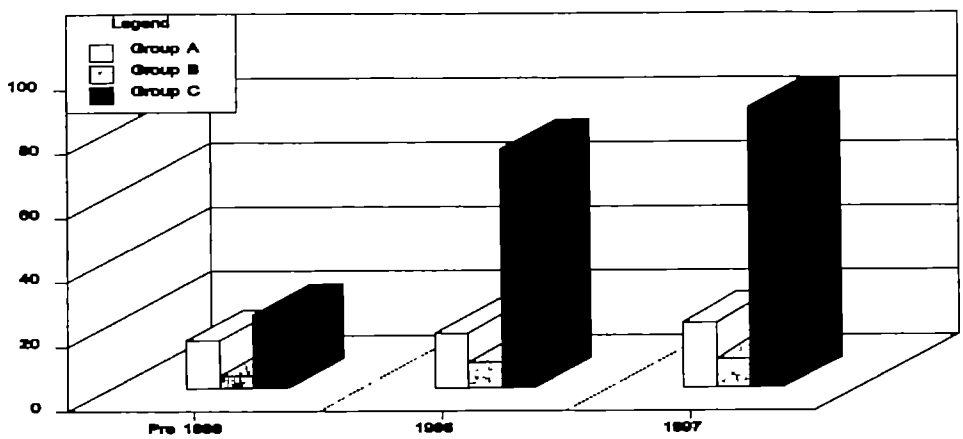
Complaints received

Universities are unlikely to expend scarce resources on measures to discourage the dissemination of defamatory material unless they are convinced that the problem is a genuine one. There are good reasons that many universities may not regard Internet defamation as a pressing concern. Universities have not been hard hit by

© The Modern Law Review Limited 1999

libel litigation arising out of the use made of their computer networks, and 37 per cent of respondents to our survey had never received any complaints that defamatory material had been accessed or distributed from their computer systems. Even at institutions that had received complaints, the number was low: most reported receiving less than three complaints a year. However, complacency may be ill-advised, as complaints are on the rise. Four universities reported ten or more complaints from January through August of 1997. Moreover, some institutions appear to be more at risk than others. While only 46 per cent of Group B universities had received complaints of defamation on their networks, 71 per cent of Group C universities had received such complaints. This latter group has also suffered a substantial rise in the number of complaints received (see Fig 1), though all groups have seen some increase.

Figure 1
Complaints by university type
(Reported to August 1997)



Universities can take solace in the fact that a substantial majority of complaints reported in our survey concern 'internal' defamation (situations where the author of the allegedly defamatory statement and the aggrieved person are members of the same institution), and most complainants are students claiming to have been defamed by fellow students. Internal complaints might be more easily resolved through a university's own disciplinary procedures than would, for example, a complaint that a leading academic affiliated with another institution had been defamed by a member of the university's staff. It should not be assumed, however, that 'internal' defamation will not lead to litigation. Even statements circulated only within a single institution can cause serious damage to a complainant, and if those statements are posted on the Internet and made available to a wider audience, the potential harm is substantial.

The reason for the steady increase in the number of defamation complaints received by universities is not clear. The provision of Internet access to staff and students has become universal only recently, and it may be that this growth in user numbers has seen a corresponding growth in misuse. It also may be that users have become more conscious of their legal rights in respect of statements which refer to them. A somewhat higher proportion of universities that offered general legal

 © The Modern Law Review Limited 1999

instruction to their users received complaints of defamation as compared to those institutions that gave no instruction in the law whatsoever,[46] and those that offered specific instruction in the law of defamation were somewhat more likely to receive complaints than those that did not.[47] The method used to deliver the training or instruction appears to be a factor: 75 per cent of universities that provided some form of verbal instruction or induction course received complaints about defamation on the computer network, while only 59 per cent of universities that merely issued a written code of practice or referred users to university regulations received complaints.

While it may be that universities offering instruction in the law have introduced the training in response to an existing problem with defamation, it would be logical to expect that the number of instances of defamation would fall once the training had been put in place, instead of showing the steady increase that the survey reveals. It seems that users are picking up enough of the anti-defamation message to render them sensitive to the need to protect their own reputations, but too little to prevent them attacking the reputations of others. Universities are thus placed in a difficult position. If an institution is sued for libel as a result of material posted on its computer network, it can rely on the defence of innocent dissemination only if it has taken reasonable care, which at the very least is likely to require that users have been given a clear indication of their responsibilities. On the other hand, by making increased efforts to educate users, a university may also be increasing the chances that a defamation action will be brought against it.

Codes of practice, training, and instruction

Publishers in the conventional media take various precautions to minimise the potential legal problems inherent in the mass dissemination of information and opinion. A newspaper or magazine publisher in the United Kingdom will typically establish a system of control designed to prevent the publication of unlawful or defamatory matter. Journalists and editors are trained in the pitfalls of defamation laws, for example, and articles are commonly screened by lawyers before they are disseminated to the public at large. Most individuals who publish on the Internet, on the other hand, probably do not consider the legal ramifications attending particular postings, and lack the appreciation of legal niceties that most journalists and editors possess. There are, in theory, many things that Internet service providers could do to educate users, ranging from the creation of codes of practice to the provision of intensive seminars on the legal aspects of publishing. Waelde and Edwards catalogued some of the options available to universities:

> How about a crash course in the laws of defamation around the world? A quick lesson on how easy it is to become sued for Internet libel? ... [A]s a practical short-term solution, some form of information available to users of the Internet about exactly what type of language may constitute a defamation – say, each time they logged into a computer, or ... each time they logged on to the Internet – might well not go amiss.[48]

All of the respondents to the survey claimed to have in place a formal code of practice relating to use of their computer networks. It is clear from the responses

46 Complaints of defamation were received by 68 per cent of institutions that offered legal instruction to users, as against 58 per cent of institutions that offered no legal instruction.

47 Eight of the twelve respondents that offered instruction in defamation law (67 per cent) had received complaints of defamation, compared with 60 per cent of those that gave no instruction.

48 Waelde and Edwards, n 1 above, 284.

© The Modern Law Review Limited 1999

that the paradigmatic code for United Kingdom universities is the JANET Acceptable Use Policy, issued by the Joint Information Systems Committee of the Higher Education Funding Councils for England, Scotland and Wales in April 1995.[49] Twelve per cent of the sample had adopted the JANET use policy without alterations or additions, and a further 56 per cent stated that they had either incorporated the JANET code into a more elaborate, institution-specific code of practice or else used it in conjunction with their own regulations. Other universities which have developed independent codes of conduct may have used the JANET policy as a template; only two institutions claimed to have a code of practice which pre-dated the JANET code.

In itself, the bare existence of a code of conduct for computer users may not be enough to show that reasonable care has been taken in relation to material published on an institution's networks. Whatever the content of the regulations governing computer use, they will have no effect if few members of a university's staff, and even fewer of its students, are familiar with them. Users need to be made aware of the content of acceptable use codes and, just as importantly, provided with some awareness of the kind of conduct that might amount to 'unacceptable' use. The JANET Acceptable Use Policy, for example, is relatively brief, stating that the network linking United Kingdom universities 'may be used for any legal activity that is in furtherance of the aims and policies' of the institutions using it, subject to the policy's provisions concerning 'unacceptable use'.[50] The provision addressing defamation is cursory: 'the creation or transmission of defamatory material' is identified as an unacceptable use,[51] a statement which hardly imparts a journalistic understanding of defamation to those to whom it is addressed. A provision such as this will have very limited effect unless users have some understanding of the kind of material likely to be regarded as defamatory.

The questionnaire asked whether institutions offered training to system users and, if so, whether any instruction was given to users in respect of their general legal responsibilities. Only one institution indicated that it offered no training of any kind, although three other institutions did not respond to these questions. The remainder offered some kind of training to both staff and students with access to computer systems. Instruction regarding the legal responsibilities of computer users was also common, with 70 per cent of respondents providing it. Legal instruction might cover intellectual property rights as well as criminal activities such as transmission of pornographic material and unauthorised access to computer systems, issues which may be of greater concern to computer system operators than defamation. Legal instruction was much more likely to be provided by the older universities in Groups A and B (80–89 per cent) than by the new universities in Group C (50 per cent). This did not appear to be strictly a question of resources; institutions with high user-to-support staff ratios were just as likely to provide legal instruction as those with lower ones.

Where legal guidance was provided to users, the method of delivery varied markedly from university to university. Institutions divided fairly evenly between those offering verbally-delivered induction courses and those which simply referred users to university regulations or a written code of practice. The older Group A universities and those with lower user-to-support staff ratios were more

49 See <http://www.ja.net/documents/use.html>. JANET is the designation of the United Kingdom's largest computer network, which links universities, other higher education institutions, and research bodies.

50 *ibid* para 8. The policy's 'unacceptable use' provisions appear in paragraph 9 of the policy.

51 *ibid* para 9.3.

© The Modern Law Review Limited 1999

likely to offer verbal instruction and induction courses. This method of delivery is more effective in making users aware of their basic legal responsibilities, even if it cannot ensure that they abide by them. Codes of practice and regulations, on the other hand, may never actually be read by users. Perhaps recognising this possibility, a few institutions asked users to sign an agreement binding them to comply with any regulations or acceptable use policy governing computer use. Signing a formal agreement may encourage a system user to read the relevant regulations, if only to discover to what he or she has agreed to be bound. Such an agreement might provide better evidence that reasonable care had been taken than would merely informing users of the existence of a code of practice. Nonetheless, it is questionable whether cursory policy statements without additional clarification and instruction are effective.

While most universities responding to the survey took steps to acquaint system users with their legal responsibilities generally, far fewer were concerned with familiarising users with the basic principles of defamation law. Less than 30 per cent overall offered this type of instruction, with Group B universities far more likely to provide such instruction (45 per cent) than institutions in Groups A or C (20 per cent and 25 per cent respectively). Availability of resources appeared to be a factor here: institutions with low user-to-support staff ratios were substantially more likely to provide users with instruction regarding defamation law. Again, methods of delivery varied. Around a third of respondents offering instruction gave it in verbal form, while the remainder either gave each user a copy of guidelines or regulations, or displayed notices in computer laboratories or on Web pages.[52] While oral instruction is probably the safest way to ensure that users are sufficiently aware of the dangers of defamation, this form of training requires the greatest use of scarce staff resources and was, not surprisingly, much more likely to be offered by those institutions with the lowest ratios of users-to-support staff. The extent of the instruction offered also varied greatly. Some respondents indicated that their code of practice merely warned users, in the manner of the JANET code, not to publish defamatory material on the network. Others gave more detailed guidelines, urging users to be certain, before publication, that facts concerning individuals or organisations are accurate and verifiable. Some also requested that, where the user is in doubt, material should be checked by a senior university official, such as the Registrar.

Disciplinary policies and procedures

A university's duty to exercise reasonable care in relation to material published on its networks may extend beyond an obligation to educate system users. At common law, distributors wishing to rely on the innocent dissemination defence needed to show they had procedures in place for removing publications from circulation if put on notice that those publications contained defamatory material.[53] Similarly, a university wishing to assert the statutory defence might be required to show that it has a clear policy for dealing with complaints concerning defamatory material

52 Only one university followed Waelde and Edwards' suggestion that information in some form should be presented to users each time they logged onto the Internet from university computers, see Waelde and Edwards, n 1 above, 284, even though this would be a quick and technically easy measure to implement

53 See eg *Vizetelly* v *Mudie's Select Library Ltd* [1900] 2 QB 170. While a library was not required to examine the entire contents of every book displayed on its shelves, it would have been required to remove a book from circulation if alerted to its defamatory contents. *Weldon* v *Times Book Club Co Ltd* (1911) 28 TLR 143.

© The Modern Law Review Limited 1999

published on its computer network. Devising such a policy is not easy. On the one hand, a policy of instituting formal disciplinary procedures in the event of complaints about a user's postings might provide the best evidence of reasonable care in the event a university is required to prove the innocent dissemination defence. On the other hand, an informal approach might be better suited to fostering an atmosphere in which the parties can be reconciled without the need for the intervention of the courts.

Of the universities that responded to the questionnaire, 65 per cent had formal procedures in place for dealing with defamation complaints, and a further 16 per cent had informal policies. Of those universities that had no policy, half had never received a complaint that defamatory material had been published on their network systems, and presumably felt no need to formulate a policy to deal with a situation that had not arisen. Less understandably, some respondents that had received complaints had no policy for dealing with such complaints. The absence of some kind of procedure in a situation where the problem is known to exist could be regarded by the courts as want of reasonable care.

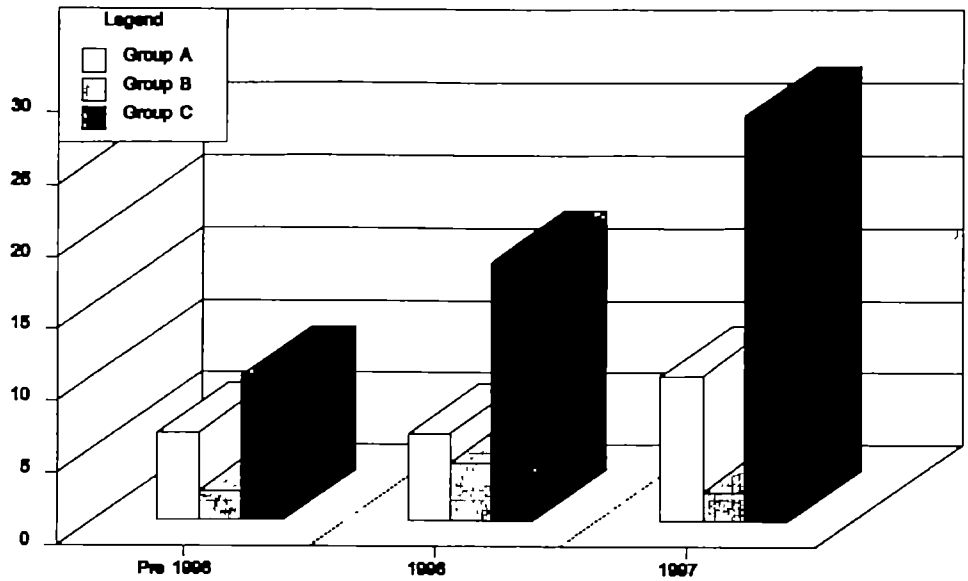
The procedures used for handling complaints vary greatly from university to university. Most commonly, incidents are investigated first by the computer services department. Then, if the complaint is considered to be justified, formal university disciplinary proceedings are invoked. Group C universities in particular tended toward the formal disciplinary approach, perhaps because the generally higher number of complaints reported by this group provoked firmer policy responses. In some institutions, complaints are handled entirely by the computer services department, either informally on a case-by-case basis or as part of a formal complaints procedure, with only the most serious cases resulting in disciplinary proceedings involving senior university officials. In others, computer services play only a minor role, with complaints being referred to the user's department or faculty. This was not a common policy, but, unsurprisingly, it is more popular in institutions that have a high ratio of users to support staff, where computer services departments might lack resources to devote to complaints of misuse. One possible drawback to this policy is that complaints might be dealt with inconsistently throughout the university, presenting problems should the institution be called upon to justify its complaints-handling policy.

Another area in which there was considerable variation among respondents concerned the willingness to withdraw access to the network from users who were the subject of complaints. Universities are generally reluctant to deny access to users unless there is clear evidence that such a step is necessary. Only around 12 per cent of respondents stated that they would immediately deny access to a user while a complaint was investigated. Some respondents were worried about the adverse effect that suspending access might have on a student's academic progress, and took the view that the power to deny access should be used sparingly, and only in serious cases after disciplinary action had been taken. If a university permits a user who has posted defamatory material on the Internet to have continued access to the network, however, it could find itself on shaky legal ground if the user continues to use the computer system to defame others. One of the factors to be taken into account in determining whether reasonable care has been taken in relation to a defamatory publication is 'the previous conduct or character of the author ... primarily responsible for publishing the statement'.[54] This factor was inherited from the common law, where a secondary publisher who distributed the

54 Defamation Act 1996, s 1(5).

© The Modern Law Review Limited 1999

Figure 2
Requests to deny access
(Reported to August 1997)



works of a primary publisher with a reputation for making libellous statements could not successfully rely on the innocent dissemination defence.[55] Arguably complaints about students or staff disseminating defamatory material will put universities on notice that those users may engage in defamatory activities in the future, and failing to revoke their privileges could be deemed unreasonable.

Much may depend on the attitude of the person defamed. A demand that the defamer be denied further access to the means of publication may be an instinctive response to the discovery that one has been libelled on the Internet, and the number of such demands is increasing[56] (see Fig 2). The survey results, however, indicate that such demands are by no means inevitable. While many institutions had received requests to deny privileges to alleged defamers, others had not, even when complaints about defamation had been received. Even if the aggrieved party initially demands that the defamer be denied access, he or she might ultimately be satisfied with less. One respondent had received a number of such demands but had not been forced to deny computer privileges to the defamers, as each complainant eventually settled for withdrawal of the material and an apology. It is doubtful, however, that an apology will always be considered sufficient vindication, particularly if the victim of the defamation is well known among his or her peers and has been seriously harmed by comments published on the Internet.

The majority of universities, whether or not they had been faced with requests to deny access, had some sort of formal or informal policy for dealing with this

55 See eg *Goldsmith* v *Sperrings Ltd* [1977] 2 All ER 566.

56 Overall, 44 per cent of respondents had been asked to deny access to one or more users, but 65 per cent of universities in Group C had received requests to deny access, as compared with fewer than one third of institutions in Groups A and B. This doubtless reflects the higher number of defamation complaints received by Group C universities, for requests to deny access appear to be increasing in proportion to the increase in complaints.

© The Modern Law Review Limited 1999

eventuality. As with procedures for dealing with defamation complaints, the new universities of Group C were most likely to have formal procedures, although this group also had the highest number of institutions with no policy at all relating to requests to deny access.[57] The formal policies of universities in Groups B and C most commonly involved decision-making at a very senior level, while in the older Group A universities, requests to deny access were usually handled entirely by the head of the computer services department.

The primary aim of any policy relating to complaints and disciplinary procedures must be to deter users from publishing defamatory comments in the first place, and this goal is unlikely to be achieved unless users are aware of the probable consequences of their actions. Most universities that had policies for dealing with defamation complaints and requests to deny user privileges published those policies to users in one form or another, though procedures relating to denial of access were less frequently published. The better resourced a computer services department, the more likely it was to publish its policies to users. The most commonly used method of publication was as part of the university regulations or some rule book, which may or may not be consulted by users. More proactive measures, such as displaying the necessary information every time a user logged into the network, or requiring users to read and sign an agreement containing details of the policy, were rarely used. Quite apart from acting as a deterrent to computer misuse, taking comprehensive steps to make users aware of the consequences of defaming someone on the Internet could help to show that reasonable care has been taken, should it become necessary to do so. It may also be important for the university's own internal disciplinary proceedings: a user who has had access suspended or other penalties imposed as a result of a complaint might well have grounds for appeal if the institution's disciplinary policy has not been made clear.

Monitoring users

One possible means by which a university might exercise care in relation to statements published via its computer network is through a policy of randomly monitoring material sent by e-mail or posted onto the World Wide Web by staff and students. Quite apart from indicating that reasonable care had been taken for the purposes of the innocent dissemination defence, such a policy would also help expose other instances of unacceptable use. While this might appear an attractive option at first glance, it gives rise to a number of potential problems. Some institutions may lack the resources to carry out continuous random monitoring. There is also the apparent paradox of the innocent dissemination defence itself: universities may be afraid of crossing the line between reasonable care and editorial control. There is probably also a reluctance to carry out what might be considered a gross invasion of privacy, particularly in respect of the monitoring of e-mail messages. Universities are also cognisant of a consideration that may be of less importance to other organisations: the need to foster and support academic freedom. The promotion of a free exchange of information and ideas does not sit comfortably with a practice of eavesdropping on student or staff communications.

57 Sixty-five per cent of universities in Group C had formal policies for dealing with requests to deny access, compared to 57 per cent of universities in Group A and 62 per cent in Group B. Twenty-nine per cent of Group C respondents had no policy for dealing with requests, compared with 21 per cent of Group A and only 13 per cent of Group B.

© The Modern Law Review Limited 1999

Whatever the reason, universities appear to be reluctant to carry out random monitoring of the users of their networks. Less than 12 per cent of respondents had any kind of policy of random monitoring. Of the few that did, less than half had a formal policy and only half publicised the policy to system users through induction courses or acceptable use policies. One respondent doubted whether random monitoring would be either legal or ethical, while another worried that if random monitoring was carried out, the university would be liable for anything that slipped through.

Universities more commonly had policies which *prohibited* the random monitoring of material sent by e-mail or posted on the Internet by users. Overall, 28 per cent of respondents claimed to forbid random monitoring. There was a very obvious split between pre-Robbins and post-Robbins universities. While 53 per cent of respondents in Group A had a formal prohibition on random monitoring, less than 12 per cent of universities in Groups B and C had similar policies. It may be that the older institutions, with a longer tradition of defending academic freedom, place more importance on users' rights to disseminate information without interference. Universities with good support staff ratios were more likely to forbid random monitoring of material transmitted on their networks, despite their greater resources for carrying out such monitoring, indicating that staffing levels do not feature as a factor in the decision of whether to monitor. In fact, of the small percentage of universities that had a monitoring policy, four-fifths were institutions with a relatively high level of users to support staff, where finding the resources to carry out monitoring might be difficult.

While the questionnaire asked only about random monitoring, some respondents noted that a particular user might be monitored in cases where misuse was suspected, even if random monitoring was prohibited.[58] One also pointed out that there was a distinction between material posted by a user on the World Wide Web and material sent by e-mail, and that monitoring the latter would be unethical. Material placed on the World Wide Web is in the public domain, available for anyone to read, and could be monitored freely without any breach of privacy.[59] E-mail messages, on the other hand, are regarded by many as private correspondence analogous to letters entrusted to the post office. Just as the post office would be protected by the innocent dissemination defence if a letter contained libellous statements,[60] some argue, so should Internet service providers who facilitate e-mail communication.[61] This notion is reinforced by the recent incorporation into UK domestic law of the right to privacy protected by the European Convention on Human Rights,[62] although universities may

58 They also indicated that general monitoring might be required in cases of system breakdown or for system security.

59 Publication on the World Wide Web might also be expected to be of most concern to universities wishing to avoid the threat of a defamation action, since this material might be expected to reach a far wider audience than most e-mail messages. There is, however, no evidence that universities are carrying out any kind of systematic monitoring of information placed on the Web by their users

60 See eg K. Norrie, *Defamation and Related Actions in Scots Law* (Edinburgh: Butterworths, 1995) 87.

61 See Gringras, n 2 above, 95 (interpreting Defamation Act 1996, s 1(3)(c)).

62 Human Rights Act 1998. Article 8 of the Convention provides that '[e]veryone has the right to respect for his private and family life, his home and his correspondence'. In *Halford* v *The United Kingdom*, *The Times* 3 July 1997, the European Court of Human Rights held that intercepting and monitoring a public employee's telephone calls made from her employer's internal telephone network without warning that her calls might be intercepted or monitored violated the employee's 'reasonable expectation' of privacy in violation of Article 8 This holding would not preclude monitoring of information placed on Websites or posted on distributed message databases, since such postings are intended for public consumption and no reasonable expectation of privacy would attach to them. But the case's reasoning could extend to e-mail communications, since users of e-mail systems arguably have a reasonable expectation that their messages are private (at least absent express warnings to the contrary).

© The Modern Law Review Limited 1999

not be legally bound to respect that right.[63] It is worth noting that many companies in the United States and the United Kingdom providing Internet access for their employees routinely monitor e-mail messages in order to ensure that confidential information is not leaked. Moreover, commercial Internet service providers in the United States monitor the postings of their subscribers and will terminate the contracts of those who violate company policies concerning the use of their facilities.[64] These activities of commercial organisations might influence a judicial determination of whether monitoring is an aspect of the reasonable care that an 'innocent' disseminator is expected to take, the principles of academic freedom notwithstanding.

Monitoring the Internet

Monitoring the individuals who use a university's computer system was widely considered problematical. But universities were less reluctant to restrict the material to which their users have access. About 55 per cent of respondents blocked access to some information available on the Internet. Of these institutions, around a third restricted access to newsgroups, many of which are arguably of little academic significance and which often contain obscene, racist or abusive messages. While the majority of universities that limited access to newsgroups did not block access to all newsgroups, a few did. Web sites known to contain pornographic material were also often restricted. The issue of blocking access to certain sites is only partly one of censorship. A small number of respondents commented that if access to newsgroups and pornographic sites was not restricted, their computer systems would not be able to cope with the volume of access demanded!

Only one respondent reported a systematic procedure for blocking access to potentially troublesome sites. In this institution, the 50 most accessed sites each week were checked for content, with objectionable sites being referred to an Information Systems Board for blocking. If a user attempted to access the blocked site, a screen message was presented which stated that access was blocked and reminded the user of the JANET Acceptable Use Policy. Users who required access to the site for legitimate academic purposes were invited to seek assistance. This approach merits consideration, although its implementation would require available staff to check and monitor suspect sites.

For institutions lacking the staff necessary to check the Internet for offensive or illegal material, software that automatically blocks access to sites with undesirable content could present a solution, but apparently not one which had much appeal for universities. Only 12 per cent of respondents used any form of blocking software, although two additional institutions were considering its use at the time of the survey. Of these few, most are using software to block 'spam' e-mail (e-mail which does not originate from a properly set up sender). The use of blocking software does present certain problems. One respondent commented that it was not technically feasible or managerially sustainable, because manpower would be required to specify and update what was to be blocked. Another asked: 'How would we censor? What rules would we apply?'. In fact, software essentially designed to reassure parents that their children are unable to access undesirable

63 Article 8 only limits the power of 'public authorit[ies]', and it is unclear (and perhaps doubtful) that universities fall within this category.

64 See K. Siver, 'Good Samaritans in Cyberspace' (1997) 23 *Rutgers Computer and Technology LJ* 1, 43.

© The Modern Law Review Limited 1999

material on a home computer is probably inappropriate for the environment of a university, which would require greater flexibility than the currently available software allows. In any event, as noted above, blocking software would have only limited effectiveness in the context of defamation.

Conclusions

In the long term, the nature and extent of the limitations placed on Internet communications will largely turn on the response of service providers to the legal and political pressures placed on them to police the Internet. The most ominous threat to freedom of speech in this new medium would be a legal regime which, intentionally or unintentionally, promotes over-zealous private censorship by these institutions. When the law fails to provide clarity and certainty in delineating the legal responsibilities of Internet service providers, the danger that these institutions will suppress or discourage lawful and socially useful communication is at its greatest. The ambiguities and inconsistencies of the innocent dissemination defence, therefore, are cause for concern.

The Defamation Act 1996 sends mixed signals to service providers regarding their responsibility for what is posted and accessed through their computer systems. On the one hand, any steps taken to exercise a measure of control over the content of Internet communications might be interpreted as inconsistent with being 'only involved' as a provider of access to the Internet, as evidence that computer users are not beyond the 'effective control' of service providers, or even that service providers exercise editorial control over the messages posted by their users. On the other hand, a *laissez faire* approach could be interpreted as a lack of reasonable care, particularly when alternatives are available that conceivably could reduce, or limit access to, abusive messages on the Internet. The frustration of the respondent who complained that 'the law is a problem in itself' is perfectly understandable.

The statutory innocent dissemination defence is vague, and vagueness has long been considered one of the greatest threats to freedom of expression.[65] The uncertainty vagueness engenders can lead to the suppression of opinions and ideas that do not cause the harms that lawmakers had sought to prevent. Thus far, universities have resisted the temptation to over-react to the potential problems caused by the ambiguities of the innocent dissemination defence. The policy responses of university computer service departments to the potential misuse of their facilities have not been uniform and sometimes seem erratic, but this is hardly surprising given the lack of clarity in the law. But regardless of differences in policy responses, most institutions have been wary of interfering with their users' rights to privacy and free expression, even when the requirements of the innocent dissemination defence arguably require such interference. Universities in the United Kingdom have not become paternalistic censors of the messages communicated over their computer systems.

In part, this simply may be the consequences of a cost-benefit analysis. Universities have received few complaints arising out of Internet defamation. The economic cost of extensive training programmes, random and non-random monitoring, and the like may simply outweigh the risk of an occasional adverse libel judgment. This might change if a university becomes entangled in an

65 *Reno* v *American Civil Liberties Union*, n 11 above, *49.

© The Modern Law Review Limited 1999

expensive libel suit, but thus far issues like pornography and copyright have been of greater concern to computer service departments than defamation.

But another consideration may be influencing the decision-making of universities in connection with the Internet. Thus far, the academy has been perhaps the greatest beneficiary of unfettered communication on the Internet. The easy dissemination and exchange of ideas on a global scale made possible by linked computer networks has revolutionised academic discourse. That discourse remains rooted in a long-standing tradition of academic freedom that institutions of higher learning will be loathe to abandon. In its formative years, the Internet came to accommodate a strongly libertarian culture hostile to governmental or institutional interference with the ability of individual users to say what they wished. Messages were to be interpreted within the cultural context in which they were posted, and a sceptical attitude rather than the threat of litigation was the primary defence against abuse.

It is unrealistic, however, to expect that this libertarian culture will survive as the Internet comes into more mainstream use. Like most new communication media, the Internet experienced a period of non-regulation, but as it has grown in influence, policy-makers have sought to exercise some control over it. Universities are not immune from this general movement away from the Internet's early libertarian ethos; the measures that have been taken to control pornography and hate speech and protect intellectual property rights are part of a broader trend that insists on the adoption of legal controls on Internet communication. It would be surprising if defamation law does not become part of this wider trend, and indeed our research indicates that complaints of defamation to university computing departments are on the rise.

In light of this, it is worth re-emphasising that much of what universities are now doing may be found to fall short of the requirements of due care. Most institutions rely on the promulgation of codes of practice containing simple exhortations against defamation rather than more thorough instruction emphasising that factual statements should be verifiable and delineating the sorts of statements that can give rise to complaints. There are dangers inherent in handling complaints in an *ad hoc* manner, and allowing users who have abused their privileges in the past to have continued access to computing facilities. A particularly controversial question – one that is thus far unanswered – is whether and when service providers will be expected to engage in random or targeted monitoring of the messages of staff and students. Ultimately, however, courts may conclude that what *most* institutions do will be the benchmark for determining whether any single institution has acted with reasonable care. If the circumscribed approach now taken by most universities comes to be regarded as 'industry practice' within the university sector, the parameters of that industry practice may go a long way in defining what reasonable care means in the context of the innocent dissemination defence.

Perhaps the best summary of the uncertainty and risks faced by universities was provided by one of the respondents to our survey, and it is appropriate to end with those words: '[Newsgroups] and the Web are both potential bombs waiting to go off. Users positively require both services for good academic reasons. At some point, an unlucky university will be in the courts or the tabloids. There are precious few controls at the moment'.

© The Modern Law Review Limited 1999

LEGISLATION

The Human Rights Act and Parliamentary Democracy

K.D. Ewing*

The Human Rights Act 1998 is the culmination of an aggressive campaign for the incorporation into domestic law of the European Convention on Human Rights, a campaign in which the judges joined forces with other political activists. Variously described as 'brilliant';[1] 'a masterly exposition of the parliamentary draftsman's art';[2] and even 'a thing of intellectual beauty',[3] the Act has also been greeted as an 'ingenious compromise' between the 'maximalists' and 'minimalists', the former supporting a judicial power to invalidate legislation, as is the case in Canada.[4] But although it is purported to reconcile in 'subtle' form the protection of human rights with the sovereignty of Parliament (a claim even more credible after an important but unsung Commons amendment), the Act also represents an unprecedented transfer of political power from the executive and legislature to the judiciary, and a fundamental re-structuring of our 'political constitution'.[5] As such it is unquestionably the most significant formal redistribution of political power in this country since 1911,[6] and perhaps since 1688 when the Bill of Rights proclaimed loudly that proceedings in Parliament ought not to be questioned or impeached in any court or any other place. In the words of Baroness Williams of Crosby, we have crossed our 'constitutional Rubicon',[7] at least to the extent that the courts may now declare a statute incompatible with Convention rights.

The enactment of the Human Rights Act 1998 was marked by an unusually high quality of debate, particularly in the House of Lords, with powerful performances by the Lord Chancellor and Lord Lester of Herne Hill. The performance of the latter on one occasion was too much for Lord Campbell of Alloway, unaccustomed it seems to cerebral rather than ceremonial contributions in their Lordships' House. Lord Campbell was heard to protest (in a manner dubbed 'offensive' by Lord Ackner)[8] about 'these escoteric lectures on law at this time of night', having 'never heard anything like this after dinner' in the 18 years he had been in the Lords.[9] He had better get used to it: the law and the lawyers are about to dominate all the debates, at all times of the day and night, before and after every meal time. Indeed in the Second Reading Debate on this Bill in the House of Lords – 'full of

* School of Law, King's College London.

1 HL Deb vol 582 col 1256 3 November 1997 (Lord Scarman).
2 *ibid* col 1234 (Lord Kingsland).
3 *ibid* col 1286 (Earl Russell).
4 HL Deb vol 583 col 1112 27 November 1997 On Canada see T. Ison, (1997) 60 MLR 499.
5 J.A.G. Griffith, 'The Political Constitution' (1979) 42 MLR 1.
6 See HL Deb vol 582 col 1234 3 November 1997 (Lord Kingsland).
7 *ibid* col 1298.
8 HL Deb vol 583 col 829 24 November 1997
9 *ibid* cols 827–828. He was also heard to complain 'why we have to be entertained for 20 – something minutes with cases from here, there and everywhere totally defeats me'.

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

immensely distinguished lawyers from all over the place'[10] – serving and retired judges alone made about one third of the speeches. Only one (Lord McCluskey, well known for his opposition) spoke against the extension of judicial power which the Act entails, with even Lord Donaldson of Lymington, a long time opponent of incorporation throughout the Thatcher/Major years, confessing that he was now persuaded of the case in favour. This is a matter to which we return.

Bringing rights home

As a democratic socialist party the Labour Party has always recognised the central importance of civil liberties, human rights and the protection of minorities. Indeed the Labour Party election manifesto in 1945 proclaimed that the Party 'stands for freedom – freedom of worship, freedom of speech, freedom of the Press', undertaking to ensure that 'we keep and enlarge these freedoms, and that we enjoy again the personal civil liberties we have, of our own free will, sacrificed to win the war'.[11] These sentiments have been expressed periodically ever since, in policy documents and in conference resolutions, it being acknowledged in 1990 that '[t]he protection and extension of individual rights has always been an essential item on the democratic socialist agenda'.[12] But notwithstanding this commitment to liberty, the Labour Party has also had an ambiguous relationship with the European Convention on Human Rights, and a reluctance to entrust the protection of rights to the judges, hardly surprising for a reforming Party which as late as 1990 saw the erosion of rights as much the responsibility of 'judicial intervention' as 'government action'.[13] Nevertheless it was the post-war Labour government which in 1951 ratified the treaty,[14] and the Labour government of Harold Wilson which in 1966 accepted the right of individual petition.[15]

But although Labour governments were prepared to go this far, the Labour Party still had difficulty with the domestication of the Convention. Private members bills in the 1980s were opposed by the Party in the Commons, while Party policy expressly repudiated the introduction of the Convention's sibling, namely a Bill of Rights. In the final report of Labour's Policy Review for the 1990s,[16] the idea of a Bill of Rights was dismissed as not providing the protection for individual rights which the Party then regarded as necessary, these relating to the right of every individual to equal treatment; the right to know how the government takes its decisions; the right to privacy; and the right to protection against the intrusive state operating through unaccountable security services. A number of reasons were given for rejecting a Bill of Rights ('or one of the variants upon that constitutional proposal which [was then] being canvassed'), including the fear that 'it could be repealed by a government with no concern for individual liberty', and the need for 'constant and detailed interpretation by the courts, with no certainty that its general provisions would protect the most vulnerable members of the community'.

10 HL Deb vol 582 col 1257 3 November 1997 (Lord Holme of Cheltenham).
11 Labour Party, *Let us Face the Future* (1945).
12 Labour Party, *The Charter of Rights* (nd) 3.
13 *ibid.*
14 See G. Marston, 'The United Kingdom's Part in the Preparation of the European Convention on Human Rights, 1950' (1993) 42 ICLQ 796
15 See A. Lester, 'UK Acceptance of the Strasbourg Jurisdiction: What Really Went on in Whitehall in 1965' [1997] PL 237
16 Labour Party, *Meet the Challenge. Make the Change* (nd) 55.

© The Modern Law Review Limited 1999

The approach of the Party – which had previously had a significant influence on democratic socialist parties elsewhere[17] – was that the protection of rights was a matter for Parliament: individual rights were to be won in Parliament (limb one) and to be secured and preserved there (limb two). How else could it be in a parliamentary democracy? To this end the Party proposed to proceed by way of 'specific legislation to provide a massive extension of individual rights and to extend democracy in Great Britain' (limb one), but sought to protect this legislation from repeal by future Parliaments by its radical programme for constitutional (and in this context parliamentary) reform (limb two). In particular it was proposed that the House of Lords would be abolished and replaced with an elected Upper House of Parliament which would have 'the power to delay repeal of legislation affecting fundamental rights for the whole life of a Parliament – thus providing an opportunity for the electorate to determine whether or not the government which proposes such measures should remain in office'.[18] This would prevent 'the swift repeal of legislation on fundamental rights by any authoritarian government which might, in the future, be elected', and would 'in effect, entrench our fundamental rights legislation'. The delaying power would apply to all legislation designated as affecting individual rights.[19]

The focus switched dramatically from the parliamentary to the judicial protection of individual liberty when John Smith proposed to modernise the constitution as part of his platform for leadership of the Labour Party in 1992.[20] In particular, he advocated the incorporation of the European Convention on Human Rights, in what represented an important break with the past. After Smith's untimely death in 1994, the commitment to incorporation was maintained, with Peter Mandelson and Roger Liddle writing in *The Blair Revolution* in a manner simultaneously hyperbolic and defensive that '[w]ith a Bill of Rights enforced by a reformed independent judiciary, the modern of age citizenship will have begun, and no one will ever be able to accuse Labour of being prepared to sacrifice individual rights on the altar of collectivist ideology'.[21] There was it seems no need any longer to distrust the judges (whose need for reform extended only to their appointment by an independent commission, as well as training and retraining) in view of the strength of their 'strictures against the excesses of the Conservative government in recent years', which were such as to 'have convinced even the most prejudiced class warrior not to question judges' independence and integrity on the basis of their social background'.[22]

Labour Party policy on incorporation was developed by Jack Straw who together with Paul Boateng published a seminal document, *Bringing Rights Home*, in December 1996.[23] This proposed the judicial enforcement of the Convention, but also a 'powerful Joint Committee of both Houses of Parliament, acting as a select committee'.[24] In 1997 the Labour Party included incorporation of the Convention in its election manifesto for the first time, proposing that 'Citizens should have statutory rights to enforce their human rights in the UK courts. We will by statute

17 See B Holmstrom, 'The Judicialisation of Politics in Sweden' (1994) 15 Int Pol Soc Rev 153.
18 Labour Party, *Looking to the Future* (nd) 42.
19 For an account of internal struggles on this issue, see L Minkin, *The Contentious Alliance* (Edinburgh: Edinburgh University Press, 1991) 477–478.
20 *The Guardian* 1 May 1992. See also *The Times* 15 April 1992.
21 P. Mandelson and R. Liddle, *The Blair Revolution. Can New Labour Deliver?* (London: Faber and Faber, 1996) 196
22 *ibid* 195.
23 J. Straw and P Boateng, *Bringing Rights Home* (Labour Party, 1996)
24 J Straw, 'Bringing Rights Home' *The Guardian* 18 December 1996.

© The Modern Law Review Limited 1999

incorporate the European Convention on Human Rights into UK law to bring these rights home and allow our people access to them in their national courts'.[25] Following the election, attention turned to the manner of incorporation, with the debate focusing sharply on whether we should adopt the Canadian or the New Zealand model as a basis for incorporation. As already noted, the Canadian Charter of Rights empowers the courts to strike down legislation,[26] while the New Zealand Bill of Rights does not. Amid press reports of divisions in the Cabinet, the Lord Chancellor announced in a public lecture that the aim was to learn from the experience of others, but 'not to be constrained by it', and that the task was to 'find a distinctively British approach for our British Parliament and British Courts'.[27]

The European Convention on Human Rights

The European Convention on Human Rights deals with a wide range of issues: these include the right to life (article 2), the right not to suffer torture or inhuman or degrading treatment (article 3), the right not to be held in servitude (article 4), the right to liberty (article 5), the right to a fair trial (article 6), protection against the retrospective application of the criminal law (article 7), the right to protection of private life (article 8), rights to freedom of conscience and religion (article 9), expression (article 10), and association and peaceful assembly (article 11), and the right to marry and found a family (article 12). This is not to say that these rights are guaranteed without limit or restriction. Although some are, many are highly qualified, including those relating to privacy, conscience and religion, expression, and association and assembly. In the case of freedom of expression the Convention permits restrictions (provided that these are 'prescribed by law' and 'necessary in a democratic society') on a number of grounds, which include the interests of national security, territorial integrity or public safety, the prevention of disorder or crime, the protection of health or morals, the protection of the reputation or the rights of others, preventing the disclosure of confidential information, and maintaining the authority and impartiality of the judiciary.[28]

Everyone whose Convention rights and freedoms are violated should by virtue of article 13 have the right to an effective remedy before a national authority, while article 14 provides that the rights and freedoms set forth in the Convention are to be secured without discrimination on any ground, and specifically on the grounds of sex, race, colour, language, religion, political or other opinion and the like. There is no free standing principle of equality, as for example in the case of the Canadian Charter, with the protection against discrimination being thus confined to the discriminatory application of Convention rights and freedoms. Derogations are permitted from some of the obligations under the Convention in time of war or other public emergency threatening the life of the nation, though 'only to the extent strictly required by the exigencies of the situation'. Additional protection is

25 Labour Party, *New Labour: Because Britain Deserves Better* (1997).

26 This is unless the legislation is protected by an express provision that it is to apply 'notwithstanding' the Charter.

27 Lord Irvine of Lairg, Keynote Address to the Conference on a Bill of Rights for the United Kingdom, University College London, 4 July 1997.

28 The ECHR lacks critical appraisal from a radical perspective But for a good general account see A.H. Robertson and A.H. Merrills, *Human Rights in Europe: A Study of the European Convention on Human Rights* (Manchester: Manchester University Press, 3rd ed, 1993). And for a valuable study of its impact in selected countries, see C Gearty (ed), *European Civil Liberties and the European Convention on Human Rights* (The Hague: Kluwer, 1997).

© The Modern Law Review Limited 1999

provided by the protocols to the Convention of which the First and the Sixth may be identified as being particularly important. The former provides by article 1 that every 'natural or legal person' is entitled to 'the peaceful enjoyment of his possessions', and is not to be deprived thereof 'except in the public interest' and subject to 'the general principles of international law'. The latter states simply that 'the death penalty shall be abolished', and that 'no one shall be condemned to such penalty or executed'.

Under the terms of the Convention, complaints may be lodged either by other High Contracting States or by individuals or organisations who claim to be victims of an alleged breach of the Convention. In the latter case a complaint could be made by an individual only if the respondent country had recognised the right of individual petition, which as we have seen has been permitted from the United Kingdom since 1966. Before the procedure was radically reformed by the 11th Protocol on 1 November 1998, complaints were lodged with the European Commission of Human Rights in Strasbourg. The responsibility of the Commission was to determine whether the complaint was admissible (with a number of conditions being imposed by the Treaty), and if so to seek to promote a 'friendly settlement' between the complainant and the respondent government. Where this failed the case could thereafter be referred by the respondent government or the Commission (but curiously not by the complainant) to the European Court of Human Rights to decide whether there had been a breach of the Convention, and if so what remedy should be provided for the victim. Where a case was not referred to the Court in this way, it would fall to be determined and disposed of by the Committee of Ministers.

The foregoing procedure was subject to criticism on a number of grounds, and was no longer suitable or effective as the number of countries ratifying the Convention continued to grow with the expansion of the Council of Europe following the end of the Cold War. Protocol 11 streamlines the procedure by creating a new European Court of Human Rights to replace the existing institutions which (in the case of the Commission and the Court) will cease to exist, or (in the case of the Council of Ministers) will no longer be responsible for adjudication. The jurisprudence of these bodies will, however, continue to be relevant (both in international and domestic law), and the new procedure will be relevant in this country notwithstanding the incorporation of Convention rights in the Human Rights Act 1998, first because applicants in the domestic courts will still have the right to proceed to Strasbourg, and secondly because not all potential breaches of Convention rights will be justiciable under domestic law.[29] Under the new procedure, the threshhold question of admissibility will be determined by a three judge committee of the Court which must act unanimously. Cases ruled admissible will thereafter be heard by a chamber of seven judges, though in exceptional cases (raising important questions of interpretation) a Grand Chamber of 17 judges will adjudicate.

Although Britain does not have the worst record of compliance under the Convention as is sometimes claimed,[30] the fact that so many complaints brought against the British government have been upheld has fuelled the campaign for the incorporation of the ECHR into domestic law, so that victims of alleged breaches of the Convention could enforce them in the domestic courts. This is indeed a strong argument in favour of the Human Rights Act: if we are prepared to accept

29 See n 66 below.

30 See A.W. Bradley and K D. Ewing, *Constitutional and Administrative Law* (Harlow: Addison, Wesley, Longman, 12th ed, 1997) 471.

© The Modern Law Review Limited 1999

the judicial enforcement of human rights, why should we not permit individuals to enforce these rights in the domestic courts?[31] By signing a Treaty which allows for a right of individual petition and by subsequently recognising the right, earlier Labour governments created a climate whereby incorporation in domestic law became inevitable, particularly following the successful incorporation of the EC Treaty. It remains the case nevertheless that there are other international treaties by which this country is bound which also provide for a complaints mechanism by those who claim that their rights have been violated. This is true for example of ILO Convention 87 and the Council of Europe's Social Charter of 1961, recently reformed with the introduction of a collective complaints procedure.[32] The logic of incorporation of the ECHR leads inexorably to an unanswerable case in favour of the incorporation of these obligations as well.[33]

The nature and extent of incorporation

The Human Rights Act 1998 does not incorporate the ECHR into domestic law in the way that the European Communities Act 1972 incorporates the EC Treaty. Rather what it does is to give effect to certain provisions of the Convention and some of its protocols by providing that these so-called 'Convention rights' are to have a defined status in English law. There is no question of the Convention rights in themselves 'becoming part of our substantive domestic law':[34] rather, certain defined provisions of the Convention enjoy a defined legal status. The terms of the Convention which are given effect to in this way by section 1 of the Act are articles 2 to 12 and 14, as well as articles 1 to 3 of the First Protocol and (following a successful amendment being moved in Committee) articles 1 and 2 of the Sixth Protocol.[35] The major omissions here are articles 1 and 13 of the Convention, the former providing that 'The High Contracting Parties shall secure to everyone within their jurisdiction the rights and freedoms defined in Section 1 of [the] Convention'; and the latter providing that 'Everyone whose rights and freedoms as set forth in this Convention are violated shall have an effective remedy before a national authority'. There was little objection to the omission of article 1, which was properly excluded on the ground that it was an inter-state obligation.[36]

---

31 See T Blair, *New Britain* (London: Fourth Estate, abridged version, 1996) 83

32 See D. Harris, 'The Collective Complaints Protocol of the European Social Charter' (Council of Europe, May 1997).

33 See further on this, K. Ewing, 'Trade unions and their "human rights" in Britain' (1998) 48 (1) *Federation News* 1. See also K D Ewing, 'The Charter and Labour' in G. Anderson (ed), *Rights and Democracy: Comparative Essays in UK/Canadian Constitutionalism* (forthcoming)

34 HL Deb vol 583 col 508 18 November 1997

35 But although effect is thus given to these different measures, in some cases it is not without limit or qualification. By virtue of section 1 (2) the different articles are to be given effect subject to any derogations or reservations. Derogations are permitted under article 15 of the Convention in time of war or other public emergency threatening the life of the nation, though there are some provisions of articles 2–4 and 7 from which even in these circumstances derogation is not permitted. This power has been exercised by the British government in respect of certain provisions of the Prevention of Terrorism (Temporary Provisions) Act 1989. The Sixth Protocol was included following a backbench amendment in the Commons. See HC Deb vol 312 cols 987–1012 20 May 1998. For the government's opposition to the amendment and the reasons for it see Michael O'Brien, *ibid* cols 1004–1005. See now Crime and Disorder Act 1998, s36. An attempt to include the Fourth Protocol (not ratified by the UK) was rejected in the Lords: HL Deb vol 583 col 492 18 November 1997 (Earl Russell).

36 HL Deb vol 583 col 467 18 November 1997 (Lord Lester of Herne Hill).

© The Modern Law Review Limited 1999

But article 13 was a different matter altogether, its absence giving rise to some concern (which in the case of Lord Ackner extended to 'suspicion' of the government's motives),[37] and an equal measure of bemusement ('a curious gap' and 'one of only three imperfections in [an] otherwise brilliant Bill').[38] For the government, however, the Lord Chancellor explained that the Bill included a scheme to provide remedies for violation of Convention rights to which it was unnecessary to add. If article 13 was to be included in the Bill, '[t]he courts would be bound to ask themselves what was intended beyond the existing scheme of remedies set out in the Bill', and it might 'lead them to fashion remedies other than the clause 8 remedies' which the government regarded as 'sufficient and clear'.[39] But in an exchange with Lord Lester of Herne Hill (in which the spectre of *Pepper* v *Hart*[40] loomed very large), the Lord Chancellor accepted that article 13 could be taken into account by the courts which were also empowered to have regard to the jurisprudence relating to article 13.[41] For his part Lord Lester regretted that scholars and others would have to read complex parliamentary debates to discover this, on the ground that it ought to be made clear on the face of the statute.[42]

In giving effect to Convention rights in this way, by section 2 the Act also directs the courts to have regard to the jurisprudence of the different enforcement and supervisory bodies in Strasbourg. For this purpose the Act refers expressly to the European Court of Human Rights, the European Commission of Human Rights, and the Committee of Ministers. As we have seen, the Commission and the Committee of Ministers will have been abolished or removed from questions of adjudication by the time the Act comes into force, though the jurisprudence of these bodies will continue to be relevant. It is to be noted, however, that the courts are required simply to take into account the jurisprudence of the Strasbourg bodies, but are not bound by it, and that the government rejected an amendment in the Lords which would have imposed the stronger obligation. The government's position was rendered no more compelling for the fact that it was supported by Lord Browne-Wilkinson who in a debate on an earlier amendment had prefaced his opposition to it by announcing rather disarmingly that he was the only person present who would have responsibility for the interpretation of the Act.[43]

This is all very curious for a number of reasons. Not least is the fact that a senior member of the judiciary should be so openly (and apparently so impatiently)[44] involved in the process of legislation, particularly with regard to a measure which extends the power of the courts directly in relation to the other branches of government. Whatever other reforms may be made to the House of Lords in the near future, there must surely be a case for excluding members of the Judicial Committee (and other senior judges) from taking part in the legislative business of

37 *ibid* col 472

38 HL Deb vol 584 col 1264 19 January 1998 (Lord Lester of Herne Hill). The other two it seems are the absence of a Human Rights Commission, and the narrow standing rules.

39 HL Deb vol 583 col 475 18 November 1997.

40 [1993] AC 593.

41 HL Deb vol 583 cols 475–477 18 November 1997.

42 He had proposed an amendment which would have clarified the issue: HL Deb vol 583 col 466 18 November 1997 On the current practice see *Rantzen* v *Mirror Group Newspapers (1986) Ltd* [1993] 4 All ER 975 and *R* v *Khan* [1996] 3 All ER 289.

43 HL Deb vol 583 col 1260 19 January 1998.

44 HL Deb vol 584 col 498 18 November 1997: 'Let us not try to protect for the time being every human right that anyone can think of' – a debate on the Fourth Protocol. See also HL Deb vol 584 col 1260 19 January 1998 (debate on Report on amendments relating to the Church).

© The Modern Law Review Limited 1999

the House. Equally curious are the reasons given by Lord Lester and Lord Browne-Wilkinson for refusing to accept the legitimacy of this constraint.[45] Essentially it seems that foreign judges are not as robust as ours,[46] and are not to be presumed always to reach decisions on the Convention which English judges would regard as fair.[47] Although not wishing to be accused of chauvinism, Lord Lester nevertheless thought that the new European Court of Human Rights will 'find itself having great difficulty in developing consistent principles of law', while according to Lord Browne-Wilkinson 'there are now a number of judges from jurisdictions which in the past at least have not been famous for their defence of human rights'.[47a] Yet if we are to incorporate the Convention should we not incorporate fully the jurisprudence of the only bodies which under the treaty are in a position to give the ECHR an authoritative interpretation?[48]

The fact remains that the Convention is meaningful only because of the principles of interpretation developed in the case law, and the decisions on points of substance, issues which are of universal application. As a text the Convention is meaningless without the jurisprudence, as is true of any other legal text: to sever the jurisprudence from the treaty is like severing the limbs from a torso. But whatever the reasons for refusing to be bound by Strasbourg jurisprudence, it should be clear that the nature and extent of judicial power under the Act is greater than may be realised. Convention rights are now free standing and autonomous (even if not 'substantive') rights of British law which the judges are empowered to develop as they wish, guided but unconstrained by the Convention jurisprudence. What this in fact confirms is that what we have incorporated is not the European Convention on Human Rights, but a number of principles which happen to be included in the Convention. We have something much closer to a distinctively British Bill of Rights than is sometimes imagined:[49] not perhaps in terms of the structure of the rights themselves;[50] but certainly in terms of their interpretation and application, with British judges having the same [un]'fetter'[ed] (the word is Lord Browne-Wilkinson's) discretion as their counterparts in Canada and elsewhere.

Statutory interpretation

The first way by which the Convention will impact upon domestic law is through the obligations in respect of statutory interpretation in section 3. This provides that that 'so far as possible to do so', both primary legislation and delegated legislation are to be read and given effect to in a way which is compatible with Convention

45 HL Deb vol 584 cols 1268–1270 19 January 1998.

46 But compare the Lord Chancellor during the Second Reading debate when he attacked the Opposition for 'railing at judges from Albania and Bulgaria sitting with other judges in the European Court to determine human rights' (HL Deb vol 582 cols 1227–8 3 November 1997).

47 As for example in cases such as *Attorney General* v *Times Newspapers Ltd* [1974] AC 273; *Attorney General* v *Guardian Newspapers Ltd* [1987] 3 All ER 316; and *X* v *Morgan Grampian (Publishers) Ltd* [1991] 1 AC 1. See now *Camelot Group* v *Centaur Ltd* [1998] 1 All ER 251

47a HL Deb vol 584 col 1269 19 January 1998

48 See also HL Deb vol 583 col 512 18 November 1997: if judges are only to take account of the jurisprudence of the Strasbourg authorities, 'we cast them adrift from their international moorings' (Lord Kingsland).

49 See also *ibid*

50 Though even here it seems now universally to be accepted that 'the text was drafted by British lawyers' HL Deb vol 582 col 1239 3 November 1997 (Lord Lester of Herne Hill)

© The Modern Law Review Limited 1999

rights. This is thought to extend the current rule of construction operated by the courts whereby they will construe existing legislation 'on the basis that Parliament intends to conform to the convention in exercising legislative power and therefore ambiguities or manifest absurdities are to be dealt with by having recourse to the convention where reasonably possible'.[51] The 'stronger command' to the judges to follow the Convention was supported in the House of Lords in order to 'reduce the mis-match' between the Convention and domestic law, but also to reduce the circumstances in which the courts may feel it necessary to declare legislation incompatible with Convention rights in accordance with section 4 of the Act (considered below).[52] The Lord Chancellor expressed the view that Parliament 'must be deemed to have intended its statutes to be compatible with the Convention' and that the courts should hold that this intention 'has not been carried successfully into effect only where it is impossible to construe a statute as having that effect'.[53]

Section 3 does, however, go some way beyond cementing the current rule of statutory interpretation. Following the House of Lords decision in *R* v *Brown*,[54] that rule appears to apply (at least in respect of the ECHR) only to legislation 'being post-Convention and ambiguous'. This meant in that case that the courts were not required (and indeed perhaps not permitted) to construe the Offences against the Person Act 1861 in a manner consistent with Convention rights in a prosecution of a number of men who had been willing participants in sado-masochistic behaviour. Section 3 of the 1998 Act provides expressly that the obligation to construe legislation to conform with Convention rights applies 'to primary legislation and subordinate legislation whenever enacted'. Parliament has also been careful to point out in section 3, however, that the duty of construction which it embraces does not 'affect the validity, continuing operation or enforcement of any incompatible primary legislation'. Nor does it affect the continuing operation or enforcement of any incompatible subordinate legislation 'if primary legislation prevents removal of the incompatibility'.

But what happens if legislation – primary or subordinate – is incompatible with Convention rights, and it is not possible to construe the former to meet the demands of the latter? The answer lies in section 4 which empowers the courts to make a declaration of incompatibility where it is not possible to interpret legislation in a manner which is consistent with Convention rights. This applies both in respect of primary and secondary legislation, though in the latter case only where the primary legislation authorising the making of the subordinate instrument 'prevents removal of the incompatibility'. So far as secondary legislation is concerned, it seems likely that this power will be infrequently invoked: what is more likely to happen is that subordinate legislation will be found incompatible with Convention rights where the parent legislation confers a wide discretion on the delegated authority. In other words the body to whom legislative authority has been delegated is not required by Parliament to exercise its power in a particular

51 HL Deb vol 583 col 534 18 November 1997 (Lord Lester of Herne Hill). This may, however, understate the nature of the existing rule of construction which is variously expressed, though on one occasion it was said by Scarman LJ to mean that the courts 'will interpret statutory language and apply common law principles, wherever possible, so as to reach a conclusion consistent with our international obligations' (*Ahmad* v *ILEA* [1977] ICR 490, 502). But compare *R* v *Home Secretary, ex parte Brind* [1990] 1 All ER 469, 477 – Convention may be used where primary legislation is 'fairly capable of bearing two or more meanings' See also *Rantzen* n 42 above.

52 HL Deb vol 583 col 534 18 November 1997.

53 *ibid* col 535.

54 [1994] 1 AC 212, especially per Lord Lowry.

© The Modern Law Review Limited 1999

way which involves a breach of Convention rights, but chooses to exercise its power in a manner which has this effect. In these cases the matter would fall not to be governed by section 4 but by section 6 (considered below), as the discretionary act of a public authority which can be challenged in the courts, directly or collaterally (as for example as a defence) by virtue of section 7.

The power to grant a declaration of incompatibility is limited to the higher courts: the House of Lords, the Privy Council, and the Courts-Martial Appeal Court. In England and Wales the power may also be exercised by the Court of Appeal and the High Court; and in Scotland by the Court of Session or the High Court of Justiciary. The duty to construe legislation to comply with Convention rights is one which in contrast is imposed on all courts. The granting of a declaration of incompatibility is discretionary even where a 'mismatch' has been found, and the government resisted an amendment that it should be mandatory.[55] If granted, a declaration of incompatibility has 'no operative or coercive effect', and 'does not prevent either party relying on, or the courts enforcing, the law in question':[56] it does not affect the validity, continuing operation or enforcement of the legislation in question; nor is it binding on the parties in the proceedings in which it is made.[57] This in itself may serve to ensure that there is little incentive for litigants to appeal where they have lost under legislation which may infringe a Convention right, but which the lower court has no authority to determine. It is in any event unclear at this stage whether there would always be grounds to appeal where a case has been decided clearly and unequivocally under legislation which may prove to breach Convention rights.

It is also unclear to what extent it may be open to a group representing the public interest (such as the Equal Opportunities Commission) to seek a declaration of incompatibility in proceedings brought directly to challenge the compatibility of a statute with the ECHR. The Act does not state in express terms that an application may be made only collaterally in the course of litigation in which the validity of the legislation arises as an issue, and it is unclear whether so much may be implied by the provision to the effect that a declaration 'is not binding on the parties to the proceedings in which it is made'. If it is open for such proceedings to be brought, questions will arise as to the rules of standing applicable in such cases: should it be on the same basis as judicial review, or is there a case for a wider provision to meet the government's desired aim that the courts should be in a position to 'develop human rights throughout society'?[58] Whatever the answer to these and other questions, section 5 does make clear that where the question of a declaration of incompatibility is raised in proceedings to which the Crown is not a party, it is entitled to notice in accordance with rules of court. Section 5 also provides that a Minister of the Crown or his or her nominee is entitled on an application made to the court, to be joined as a party to the proceedings. An application may be made at any time in the course of the proceedings.

55 HL Deb vol 583 cols 545–548 18 November 1997.

56 *ibid* col 546.

57 It is also worthy of note that the government intends that the 1998 Act will not impliedly repeal any earlier inconsistent legislation: the declaration of incompatibility applies here too, and is intended as the only device for dealing with conflict between the 1998 Act and earlier legislation. This caused some bemusement in the Lords

58 HL Deb vol 582 col 1228 3 November 1997.

© The Modern Law Review Limited 1999

Liability of public authorities

The second way by which the Act gives effect to the Convention is by imposing an obligation on public authorities to comply with Convention rights, an obligation which is directly enforceable in the courts. Section 6 provides that it is unlawful for a public authority to act in a way which is incompatible with one or more of the Convention rights, with a 'public authority' being widely defined to include a court; a tribunal which exercises functions in relation to legal proceedings; and any person certain of whose functions are functions of a public nature. The definition is fluid and open ended, the government taking a policy decision to 'avoid a list',[59] and declining an invitation to identify precisely to which bodies the Convention rights should apply. But the intention clearly is that they should apply to central government (including executive agencies), local government, the police, immigration officers, and the prison service, as well as to others.[60] One of the most difficult questions under the Act relates to the inclusion of courts and tribunals within the definition of a public authority. The precise meaning and implications of this are unclear, though it seems that it is not designed simply to ensure that the courts comply with Convention rights on 'such issues as contempt in the face of the court, the exclusion of the public from proceedings, the toleration of secret witnesses, the operation of public interest immunity certificates and so on'.[61]

According to the Lord Chancellor, the government 'believe that it is right as a matter of principle for the courts to have the duty of acting compatibly with the Convention not only in cases involving other public authorities but also in developing the common law in deciding cases between citizens'.[62] But what does this mean, given the government's belief that 'full horizontal effect' would be 'a step too far in a Bill which ... is designed to allow the Convention rights to be invoked in this country by people who would have already a case in Strasbourg',[63] and given also the rejection by the Lord Chancellor of the view that the courts will be required to legislate by way of judicial decision 'whenever a law cannot be found either in the statute book or as a rule of common law to protect a convention right'?[64] The answer it seems is that Convention rights may be relied upon in litigation between private parties, but cannot themselves be the basis of a cause of action.[65] So although a worker dismissed for a reason incompatible with the Convention may not sue his or her employer for breach of a Convention right, the worker in question may be able to sue for wrongful dismissal, claiming that a dismissal for a reason incompatible with a Convention right is wrongful.[66] But it is

59 HL Deb vol 583 col 796 18 November 1997

60 *Rights Brought Home: The Human Rights Bill* Cm 3782 (1997) para 2.2

61 C.A. Gearty, 'Will Incorporation Change Anything?' *Human Rights and Scots Law* (University of Edinburgh, 26 June 1998).

62 HL Deb vol 583 col 783 24 November 1997.

63 *ibid* col 798

64 *ibid* col 785

65 1998 Act, s 7. See generally on the question of horizontality, M. Hunt, 'The "Horizontal Effect" of the Human Rights Act' [1998] PL 423.

66 This is nevertheless less than full repatriation of the Convention rights, in the sense that it is not possible to bring an action in the British courts to cover all the situations in which a complaint may be made by an individual to Strasbourg. it is difficult for example to see how a case such as *Stedman* v *UK* (1997) 23 EHRR CD 168 (in which an employee of a private employer complained that she was dismissed for refusing to work on Sundays) could begin life in the English courts. On the application of the Act to employment law, see K.D. Ewing, 'The Human Rights Act and Labour Law' (1998) 27 ILJ 275, which deals also with the question of how far Convention rights apply to private action.

© The Modern Law Review Limited 1999

too early to say whether the Act will have quite the spectacular effect on the common law that some of the more extravagant claims currently suggest.

Apart from courts and tribunals, a public authority is defined only specifically to include bodies which are not exclusively public bodies, but some of whose activities are of a public nature. Here section 6 provides that such a body is a public authority, except in relation to a particular act where 'the nature of the act is private' (section 6 (3) (c)). An example of a body which falls within this category is Railtrack 'because some of its functions, for example its functions in relation to safety on the railway, qualify it as a public authority'. But the Lord Chancellor was of the view that its work as a private property developer would 'no doubt be held by the courts to be of a private nature',[67] though it is perhaps surprising that a statutory authority should be regarded as having mixed functions in this way, which seems perfectly appropriate for private companies carrying out public functions as government contractors.[68] We nevertheless have the complex position that Convention rights are (a) directly enforceable against some bodies in respect of all of their activities (even when acting as employers for example),[69] (b) directly enforceable against some bodies in respect of some but not all of their activities (public but not private),[70] and (c) not directly enforceable at all against others (exclusively private). The critical point then will be to determine what is a function of a 'public' nature, a matter on which the Act gives no guidance whatever.

But although the definition of public authority is widely drawn, the right to bring an action is limited to victims of the alleged unlawful action, defined as someone who would be a victim for the purposes of article 34 of the Convention. Proceedings for breach of a Convention right by a public authority may be brought in the appropriate court or tribunal, and may be brought in an application for judicial review.[71] Where an action proceeds by way of judicial review, the applicant is to be taken to have a sufficient interest only if he or she is or would be a victim of the act complained of. This reflection of the terms of the Treaty in the statute attracted considerable criticism in Parliament, and an amendment was moved unsuccessfully to remove the restriction that only victims can complain, and thereby apply the current rules of standing in judicial review proceedings.[72] The government resisted on the ground that the intention is to allow people to proceed in the domestic courts on the same basis as they can proceed to Strasbourg, pointing out that interest groups will still be able to provide assistance. An action must be brought within a year of the conduct complained of, 'subject to any rule imposing a stricter time in relation to the procedure in question', a measure which presumably would require judicial review proceedings to be brought within three months.

---

67 HL Deb vol 583 col 796 24 November 1997.

68 As in the case of private security companies in relation to the management of prisons: *ibid* col 811

69 Compare the position in administrative law where public bodies are subject to review only in respect of their public activities. *See R* v *East Berks Health Authority, ex parte Walsh* [1985] QB 152.

70 Apart from those mentioned, others include the churches and the Jockey Club: HC Deb vol 312 col 1019 20 May 1998 (The Home Secretary, Mr Jack Straw). But judicial review proceedings under Order 53 will not always be available against such bodies.

71 It is also provided expressly that a person may rely on the Convention rights in any legal proceedings (s 7(1)(b)) Take for example the *Sunday Times* case [1974] AC 273 where an injunction was granted to restrain a publication in contempt of court, or the *Spycatcher* case [1987] 3 All ER 316 where an injunction was continued to restrain a publication in breach of confidence. In both cases the injunctions were found to be in breach of article 10 of the Convention, and if granted after the Act comes into force would mean that the courts themselves (as public authorities) would be acting unlawfully by virtue of conduct (issuing an injunction) which was incompatible with a Convention right (article 10). See also s 12, discussed below.

72 HL Deb vol 583 col 823 24 November 1997.

© The Modern Law Review Limited 1999

If a complaint about the violation of Convention rights is upheld, the court in question is empowered to 'grant such relief or remedy, or make such order, within its powers as it considers just and appropriate' (section 8). In judicial review proceedings, which seem destined to be the normal route to a remedy in this field, it means that the whole range of Order 53 remedies will be available at the discretion of the court. But damages are only to be awarded where the court is 'satisfied' that the award is 'necessary' to 'afford just satisfaction' to the person in whose favour it is made; and only after the court has taken into account any other relief or remedy granted or order made.[73] However, in determining whether to make an award of damages, and in determining the amount of any award, the court should have regard to the principles applied by the European Court of Human Rights in relation to compensation awards under the Convention, though again the courts are not bound to follow these principles. Where the public authority is a court or tribunal, damages are not to be awarded for a 'judicial act done in good faith' (section 9), except to compensate an individual to the extent required by article 5(5) of the Convention which provides that 'everyone who has been the victim of arrest or detention in contravention of the provisions of [article 5] shall have an enforceable right to compensation'.[74]

Remedial action

As we have seen, section 4 of the Act permits the courts to declare an Act of Parliament to be incompatible with Convention rights, though they are required still to enforce and apply it until such time as the legislation has been amended.[75] The government considered but rejected the option of giving the courts the power to set aside an Act of Parliament believed to be incompatible with Convention rights, 'because of the importance the Government attaches to Parliamentary sovereignty'. The White Paper continues in powerful and convincing terms, as follows:

> In this context, Parliamentary sovereignty means that Parliament is competent to make any law on any matter of its choosing and no court may question the validity of any Act that it passes. In enacting legislation, Parliament is making decisions about important matters of public policy. The authority to make those decisions derives from a democratic mandate because they are elected, accountable and representative.

Concern was expressed that to permit the courts to set aside Acts of Parliament 'would confer on the judiciary a general power over the decisions of Parliament', and 'would be likely on occasions to draw the judiciary into serious conflict with Parliament'. Crucially there 'is no evidence to suggest that they desire this power, nor that the public wish them to have it'.[76]

---

73 But it is not intended that damages can be awarded by the Crown Court where in the course of a trial the court finds there to have been a violation of a person's rights It will be necessary to go to the civil courts: HL Deb vol 583 col 855 24 November 1997.

74 *ibid* col 856.

75 There is, however, no obligation on the part of the government to bring forward amending legislation, and the failure to do so is not judicially reviewable (s 6 (6)). It was argued that to permit judicial review would be 'inconsistent with a fundamental precept of our constitutional arrangements; namely, that the courts do not interfere with the proceedings of Parliament': HL Deb vol 583 col 815 24 November 1997. But this does not seem very convincing after *R* v *Home Secretary, ex parte FBU* [1995] 2 AC 513, or in light of the fact that the decision to introduce a remedial order under section 10 would appear to be judicially reviewable, on which see this section below.

76 *Rights Brought Home* n 60 above para 2.13.

© The Modern Law Review Limited 1999

So why give the courts the power to rule on the compatibility of an Act of Parliament with Convention rights? How can this be reconciled with the authority of the mandate? The answer is that it is open to the government to decide how to deal with the decision of the courts, and to refuse to take steps to remedy the incompatibility if it deems it appropriate to do so. This after all is the position currently with decisions of the European Court of Human Rights which do not change domestic law but require amending legislation, which the government cannot be compelled to introduce and Parliament cannot be compelled to pass. But although these are weighty and logical considerations, we should be careful about distinguishing form from substance, principle from practice. As a matter of constitutional legality, Parliament may well be sovereign, but as a matter of constitutional practice it has transferred significant power to the judiciary. The cat was let out of the bag by Lord Borrie during the Second Reading debate in the Lords when he said in his speech that as a matter of 'political reality', 'while historically the courts have sought to carry out the will of Parliament, in the field of human rights Parliament will carry out the will of the courts', confident that 'government and Parliament will faithfully implement any declaratory judgment made by the High Court'.[77]

Yet while the judges may now be in the driving seat on human rights issues, we should be careful about over-stating the point. It is quite clear from the Commons debates that the government intends that Parliament should still hold the key, and that a declaration of incompatibility will not necessarily lead to legislation in all cases. At the Report stage, the Home Secretary cited the example of abortion, stating that it was 'possible to conceive that, sometime in the future, a particularly composed Judicial Committee of the House of Lords reaches the view that provision for abortion in either the United Kingdom or part of the United Kingdom is incompatible with one or another article of the Convention'.[78] He acknowledged that such a decision would 'create very great controversy and, in some quarters, considerable social anxiety', and thought that most governments would want to 'continue with the existing abortion legislation'.[79] Apart from measures dealing with social issues such as this, other examples of legislation which might well survive a declaration of incompatibility include provisions dealing with emergencies (such as parts of the Criminal Justice (Terrorism and Conspiracy) Act 1998), as well as provisions of constitutional significance, such as the Representation of the People Act 1983, section 75 of which was found to be in breach of article 10 of the Convention by the Strasbourg court.[80]

Apart from thus giving courts the power to question statutes, the government's commitment to parliamentary sovereignty was tested also by its initial proposal for dealing with declarations of incompatibility (which would apply also to decisions of the European Court of Human Rights, including decisions to which this country was not a party). Thus because of the volume of parliamentary business and the 'protracted' nature of the legislative process, a fast track procedure would be introduced whereby the government would be empowered to amend primary legislation by order. So although the courts would not be empowered to strike down legislation, they would effectively be doing so indirectly: given the fact that the government would 'almost certainly' always want to change the law,[81]

77 HL Deb vol 582 cols 1275–1276 3 November 1997.
78 HC Deb vol 317 col 1301 (The Home Secretary, Mr Jack Straw)
79 *ibid.*
80 *Bowman v UK* (1998) 26 EHRR 1
81 *Rights Brought Home* n 60 above para 2.10

© The Modern Law Review Limited 1999

ministers would become simply the medium for giving effect to the decisions of the courts, reduced to little more than judges' runners. The power which it was proposed to confer on ministers was extraordinary, in the sense that it would not be confined to a power to amend legislation to the extent necessary to deal with the incompatibility, but could include 'such incidental, supplemental, consequential and transitional provision as the person making [the order] considers appropriate'. In some cases it was proposed that the order could be made without parliamentray approval, though this would have to be secured *ex post facto*.

Although these proposals were warmly supported in some quarters, they were strongly criticised as 'unnecessary and unconstitutional'[82] in others, with concern being expressed about 'Henry VIII trampling through the statute book'.[83] The offending clause (10) was amended at the committee stage in the Commons, in perhaps the most significant victory in the parliamentary history of this particular Bill. The result is to restore the principle that primary legislation should be amended or repealed only by primary legislation, and to this extent the Act gives rather greater weight to the government's commitment to parliamentary sovereignty then had originally been proposed, by restoring Parliament's authority in the legislative process. As enacted section 10 empowers a minister to take remedial action to amend primary legislation by order only where there are 'compelling reasons' for doing so (a judgment which will doubtless be the subject of judicial review),[84] and only to the extent necessary to remove the incompatibility (though to confuse matters a remedial order may still relate to legislation other than that which contains the incompatible provision).[85] The principal route for responding to a declaration of incompatibility is thus to be legislation, it being the responsibility of Parliament in the normal way to determine whether and how far legislation is to be amended.

The press and the churches: special pleading; special privileges

In the course of the debates in Parliament and beyond, the government was found to stumble over only two issues; these relating to the application of the Convention to the press on the one hand, and the churches on the other. Perhaps predictably neither institution was happy about the possible implications of the Convention, with the former concerned that it might lead to the introduction through the back door of a legal right to privacy. The newspaper industry found itself in an awkward position as an institution prepared to exploit human rights considerations for corporate commercial purposes, while simultaneously appearing unwilling to respect the human rights of others. The concerns of the churches were rather different when it was realised that (unlike the press) they are public bodies (at least in some of their activities) and as such bound to respect Convention rights. Here too special privileges were claimed by organisations which are rhetorically committed to respect the human rights of others, but concerned that they might be required to show the same respect themselves, most notably in connection with same sex marriages, and the adoption of children by same

82 HL Deb vol 583 col 541 18 November 1997 (Lord Campbell of Alloway).

83 HL Deb vol 583 col 1141 27 November 1997 (Lord Simon of Glaisdale). See also Select Committee on Delegated Powers and Deregulation, 6th Report (HL 32, 1997/98) paras 23 and 27.

84 It is only failures to bring forward legislative or remedial orders which are protected by s 6(6). The decision to by-pass normal parliamentary procedure for 'compelling' reasons is one which ought to be justified

85 And to the extent that it now applies to decisions of the European Court of Human Rights, it does so only in relation to proceedings involving the United Kingdom.

© The Modern Law Review Limited 1999

sex couples. A number of amendments making special provision for the Church were carried in the House of Lords, and together with the press, the government accepted the need for some protection of their special status.

Press concern that the Act would lead to the introduction of a privacy law ('despite the Government's stated opposition to one')[86] was expressed in the House of Lords by Lord Wakeham. As chairman of the Press Complaints Commission (PCC), he was also concerned that the Commission would be regarded as a public authority, with the result that it would be bound to comply with Convention rights. The initial response of the Lord Chancellor was denial, but he changed his mind following the publication of counsel's opinion received by the PCC.[87] This clearly alarmed Lord Wakeham who tabled an amendment to the Bill which as a result would not have applied to courts or tribunals if 'the parties to the proceedings before [them] do not include any public authority'. Lord Wakeham's principal concern was that the system of self regulation would be fatally undermined if the PCC were to be regarded as a public authority and subject to scrutiny by the courts, protesting that his 'task of seeking to resolve differences, to obtain a public apology where appropriate or, if necessary, to deliver a reprimand to an erring editor would no longer be a practical proposition because the courts would be able to intervene after [his] work was finished'.[88] Otherwise an additional concern was that 'in privacy matters newspapers would be subject to interim as well as final injunctions under its terms'.[89]

These fears were dismissed by the government. On the right to privacy, the Lord Chancellor claimed that such a right is about to emerge anyway, because the judges are 'pen-poised, regardless of incorporation of the convention, to develop a right to privacy to be protected by the common law'.[90] Indeed the Lord Chancellor reinforced the point in the following terms:

> In my opinion, the court is not obliged to remedy the failure by legislating via the common law either where a convention right is infringed by incompatible legislation or where, because of the absence of legislation – say, privacy legislation – a convention right is left unprotected. In my view, the courts may not act as legislators and grant new remedies for infringement of convention rights unless the common law itself enables them to develop new rights or remedies.[91]

So far as the obligation of the PCC is concerned, the Lord Chancellor was equally unmoved, claiming that the courts would have to intervene 'only if self-regulation did not adequately secure compliance with the Convention', and that the message for the press was plain: strengthen self-regulation and strengthen the PCC. But despite its apparent willingness to yield, the government did consult with Lord Wakeham and the newspaper industry to deal with their concerns.[92]

The result of these consultations was the drafting ('in consultation with Lord Wakeham and representatives of the national and regional press')[93] of a new clause which was given a 'warm welcome' by the industry.[94] What is now section 12

86 HL Deb vol 583 col 1771 (Lord Wakeham).
87 For a critique of the government's position see the contribution of Mr Norman Fowler, HC Deb vol 314 col 400 *et seq* 17 June 1998.
88 HL Deb vol 583 cols 772–773 24 November 1997
89 col 773.
90 *ibid* col 784.
91 *ibid* col 785
92 HC Deb vol 315 col 541 2 July 1998 (The Home Secretary, Mr Jack Straw).
93 *ibid.*
94 *ibid.*

© The Modern Law Review Limited 1999

applies in any case where a court is considering whether to grant any relief which might affect the exercise of the Convention right to freedom of expression, including, for example, 'an injunction restraining a threatened breach of confidence' (as in *Spycatcher*).[95] Where section 12 does apply, no relief is to be provided in *ex parte* proceedings unless the court is satisfied that the applicant has taken 'all practicable steps' to notify the respondent, or if 'there are compelling reasons why the respondent should not be notified' (section 12 (2)). The latter measure is designed to deal with 'a case raising issues of national security where the mere knowledge that an injunction was being sought might cause the respondent to publish the material immediately'.[96] This restriction on *ex parte* injunctions is not unprecedented,[97] any more than the requirement of section 12 (3) which – addressing the rules established in *American Cyanamid Co* v *Ethicon Ltd*[98] – directs the courts not to restrain a publication pending the trial of an action (as in the *Spycatcher case*),[99] unless satisfied that 'the applicant is likely to establish that publication should not be allowed'.[100]

The foregoing provisions of section 12 are likely to be confined in practice to interlocutory proceedings. The third obligation on the courts is, however, likely in general to be of wider application, though whether it adds anything to an obligation which would almost certainly exist anyway is seriously open to question. But it does appear to have satisfied the press.[101] This is the duty in section 12 (4) to have regard to the importance of the Convention right to freedom of expression in cases where the granting of relief would affect its exercise. Section 12 (4) also provides that in cases which relate to journalistic, literary or artistic material, the court must have regard to the extent to which (a) the material has or is about to become available to the public, or (b) it would be in the public interest for it to be published. In these cases the court must also have regard to any relevant privacy code, so that the fact that a newspaper has complied with the terms of the PCC code will be weighed in its favour by the courts.[102] But although this gives an important legal status to the industry's own procedures, the courts are not bound by it. So the extent to which a newspaper has complied with the PCC Code may not be a decisive factor if the courts take the view in any case that the Code itself falls short of acceptable standards.

---

95 *ibid* col 535 (The Home Secretary, Mr Jack Straw). It could also include relief 'relating to criminal proceedings – which affect the article 10 right' (such as a prosecution under the Official Secrets Acts 1911–89).

96 *ibid*, col 536 (The Home Secretary, Mr Jack Straw). It is not anticipated that this qualification would be often used, and in the past such applications have been rare (*ibid*).

97 See for example Trade Union and Labour Relations (Consolidation) Act 1992, s 221. But see *Barretts & Baird (Wholesale) Ltd* v *IPCS* [1987] IRLR 3. See S. Fredman, (1987) 103 LQR 176.

98 [1975] AC 396 But see *Holley* v *Smith* [1998] 1 All ER 853.

99 *AG* v *Guardian Newspapers Ltd* [1987] 3 All ER 316.

100 1998 Act, s 12 (3). This latter provision is a recognition of the fact acknowledged by the European Court of Human Rights that 'news is a perishable commodity and to delay its publication, even for a short period, will deprive it of all its value and interest' (*The Observer and The Guardian* v *UK* (1991) 14 EHRR 153 191). For a similar provision, see Trade Union and Labour Relations (Consolidation) Act 1992, s 221 and *NWL Ltd* v *Woods* [1979] 1 WLR 1004.

101 The churches arguably got more by way of concession, with section 13 providing that if a court's determination of any question arising under the Act might affect the exercise by a religious organisation of the Convention right to freedom of thought, conscience and religion, it must have regard to the importance of that right.

102 HC Deb vol 315 cols 538–539 2 July 1998.

© The Modern Law Review Limited 1999

Parliamentary procedure and parliamentary scrutiny

The attention so far has focused on the role of the courts and the relationship between the courts and Parliament. But the Act also addresses questions of parliamentary procedure, in a manner which follows the example of the Canadian Bill of Rights 1960 and the New Zealand Bill of Rights 1990 (though neither has been expressly acknowledged by the government). In Canada section 3 of the Bill of Rights imposes a duty on the Minister of Justice to examine every Bill introduced in or presented to the House of Commons,[103] 'in order to ascertain' whether any of its provisions are 'inconsistent with the purposes and provisions' of the Bill of Rights and 'to report any such inconsistency to the House of Commons at the first convenient opportunity'.[104] A similar duty is in effect imposed on the Attorney General in New Zealand in the case of any Bill introduced into the House of Representatives, the Attorney being required to bring to the attention of the House 'any provision in the Bill that appears to be inconsistent with any of the rights and freedoms' contained in the Bill of Rights. In the case of government bills this must be done on the introduction of the bill; but in the case of other bills, it should be done as soon as practicable thereafter.

It is already the case in this country that the Ministerial Code provides that ministers must take into account obligations arising under the ECHR in preparing business for the Cabinet.[105] The government wishes to take further the 'pre-legislative scrutiny of all new policy measures'[106] in order to ensure that the legislation which it places before Parliament in the normal way is compatible with Convention rights. Consequently section 19 provides that the Minister 'in charge of a Bill in either House of Parliament' must make a statement to the effect that in his view its provisions are compatible with Convention rights, the government intending that such statements should be included alongside the Explanatory and Financial Memorandum which accompanies a Bill when it is introduced. But the government was not prepared to accept a Liberal Democrat amendment which would require the statement of compatibility to be accompanied by written reasons for the government's view why the Bill complied with the Convention. This was a response to the concern that some ministers 'rather vaguely wave their hands in the air and make statements to the effect of this or that with no great precision, no great clarity and no great helpfulness to the House'.[107]

There may of course be cases where it is not possible for the minister to make a declaration of compatibility, because for example the government feels the need to proceed with legislation without being fully aware of its implications for Convention rights. There may also be very rare cases where the government feels the need to legislate regardless of whether there is any violation of Convention rights. In these cases the minister must make a statement to the effect that although he or she is unable to make a statement of compatibility, the government nevertheless wishes the House to proceed with the Bill. Where a minister is unable to make a statement of compatibility 'that will be a very early signal to Parliament that the possible human rights implications of the Bill will need and will receive

103 It also applies to draft regulations.
104 For comment see W.S. Tarnopolsky, *The Canadian Bill of Rights* (Toronto: Macmillan, 2nd revised ed, 1975) 125–128.
105 *Ministerial Code: A Code of Conduct and Guidance on Procedure for Ministers* (Cabinet Office, 1997).
106 HL Deb vol 583 col 1163 27 November 1997 (Lord Irvine of Lairg).
107 *ibid* col 1162.

© The Modern Law Review Limited 1999

very careful consideration'.[108] Whether this is enough is of course open to question, given the tight grip that most governments now have on Parliament, and given the proposals of the Labour Party less than 10 years ago which anticipated the possibility that Parliament might be empowered to delay a Bill on the ground of a potential human rights violation for the life-time of a Parliament. In fairness, however, the government is at least contemplating the possibility of a more formal role for Parliament on human rights issues specifically beyond the obligation of ministers to report on the question of compatibility or otherwise.

The idea of a Parliamentary Committee on Human Rights is considered in both the Labour Party document *Bringing Rights Home* and in the White Paper *Rights Brought Home*. In the former it is recognised that 'Parliament itself should play a leading role in protecting the rights which are at the heart of a parliamentary democracy';[109] and in the latter it is claimed that the manner in which this is done is a matter for Parliament, though the government's view is that it could best be done by establishing a new parliamentary committee with responsibility for human rights.[110] This would not require any legislation (though it would require new Standing Orders), and it would apparently not require any change to parliamentary procedure, which suggests that the government is not contemplating a human rights committee with a role in the legislative process, unlike similarly conceived committees in other jurisdictions,[111] and indeed unlike the Joint Committee on Statutory Instruments in this country which has power to identify possible ultra vires measures.[112] In the absence of a legislative function, it is anticipated that the new Committee might 'conduct enquiries on a range of human rights issues relating to the Convention, and produce reports so as to assist the Government and Parliament in deciding what action to take'.[113] It might also want to range more widely and examine issues relating to other international obligations.

The Lord Chancellor returned to this question during the Second Reading debate in the Lords where it was now acknowledged that it was 'not clear what functions such a Committee would perform',[114] though it was contemplated that the Committee could, for example, not only keep the protection of human rights under review, but also be in the forefront of public education and consultation on human rights. He also anticipated that it could 'receive written submissions and hold public hearings at a number of locations', and as such be in the vanguard for promoting 'a human rights culture across the country'.[115] If these are to be the responsibilities of a parliamentary committee (and they would be more commonly associated with a Human Rights Commission of a kind which the government has rejected at least for the time being),[116] they should be supplementary to what ought to be the principal responsibilities of such a committee, that is to say the legislative screening and the investigation of human rights abuses by government departments (a task which would not be rendered

108 *ibid* col 1163.
109 n 23 above
110 n 60 above para 3 6.
111 D. Kinley, *The European Convention on Human Rights: Compliance Without Incorporation* (Aldershot Dartmouth, 1993).
112 Bradley and Ewing, n 30 above 727.
113 n 60 above para 3.7
114 HL Deb vol 582 col 1234 3 November 1997.
115 *ibid.*
116 See for example the powers of the Equal Opportunities Commission and the Commission for Racial Equality in the Sex Discrimination Act 1975, ss 53–56 and Race Relations Act 1976, ss 43–46 respectively. These bodies obviously have other functions as well.

© The Modern Law Review Limited 1999

impracticable by the availability of judicial review). If the government is prepared to surrender power to the courts, it ought at least to be willing to contemplate a more formal prophylactic role for Parliament in the enactment of legislation.[117]

Conclusion

The Human Rights Act represents a fundamental shift in the position of the Labour Party which less than ten years ago was still deeply suspicious about the contribution of the judges to civil liberties, and suspicious also of the idea that the courts should be in the driving seat on controversial political issues. Now ministers can refer to no less a figure than Lord Denning for the proposition that 'we have to trust someone, so why not trust the judges',[118] and attack those who question whether so much power should be given to the courts as 'ageing Marxist[s] decrying the establishment's conspiracy against the proletariat'.[119] But in a sense the Human Rights Act is a metaphor for a much deeper transformation of the Labour Party, and indeed the principles of democratic socialism to which the Party is constitutionally committed by its revised clause IV.[120] Although it is true that the principles of democratic socialism are flexible and evolutionary, it would nevertheless have been difficult historically to reconcile with these principles the quasi-entrenchment of liberal icons in the shape of the ECHR,[121] particularly when secured in a manner which allows the courts – at the suit of legal as well as natural persons – to question Acts of Parliament.[122]

But for all that, the inevitable incorporation of the Convention has been secured in a manner which subordinates Convention rights to constitutional principle and democratic tradition. We should not diminish the importance of the fact that the Human Rights Act does not give the courts the power to strike down legislation. Nor should we diminish the reasons why the government withheld this power from the courts, grounded as they were in a desire to ensure that 'no court in this land, not even the Judicial Committee of the House of Lords, can place itself in the position of sovereignty over the High Court of the elected Parliament'.[123] The government appeared particularly concerned to preserve 'the fundamental position established in our constitution: the sovereignty of Parliament', said to be 'one of the profound strengths of our system'.[124] This is a refreshing position to have adopted, if only because the sovereignty of Parliament is something which should not readily be conceded, acting as it does as a constitutional and legal principle

117 A weak parliamentary committee would be a wasted opportunity to do something worthwhile to restore the authority of Parliament.

118 Mr Michael O'Brien, HC Deb vol 306 col 857 16 February 1998, citing Lord Denning, *What Next in the Law* (London: Butterworths 1982) 330. 'Someone must be trusted. Let it be the judges'.

119 *ibid*, col 858.

120 Labour Party Rule-Book, *National Rules and Procedures of the Labour Party*, clause IV.

121 On the liberal project to constrain democracy, see C.B. Macpherson, *The Life and Times of Liberal Democracy* (Oxford: OUP, 1977). For a pithy statement of the liberal position, see K. Loewenstein, 'Militant Democracy and Fundamental Rights II' (1937) Am Pol Sc Rev 638, 658: 'democracy has to be redefined. It should be . . the application of disciplined authority, by liberal minded men, for the ultimate ends of liberal government: human dignity and freedom'.

122 See Clifford Chance, *Incorporation of the European Convention on Human Rights: The Implications for Business* (1997). Recent examples of corporate attempts to use the Convention include *National and Provident Building Society* v *UK* (1997) 25 EHRR 127.

123 HC Deb vol 317 col 1300 (The Home Secretary, Mr Jack Straw)

124 *ibid.*

© The Modern Law Review Limited 1999

which has metamorphosed in a dynamic constitution, to give effect to one of the first political principles of democratic self-government. This is the principle of popular sovereignty whereby the elected and accountable representatives of the people should be empowered by the authority of a mandate to give effect without restraint to what used to be called the General Will.

This is not to deny that the Human Rights Act confers a significant political power on the courts, or indeed that it enables the courts formally to set the agenda on human rights questions. Nor is it to deny that in the 'overwhelming majority of cases', a declaration of incompatibility will be accepted by the government and Parliament of the day, thereby ensuring that the elected branch will in practice defer to the unelected branch.[125] But the government appears genuinely to anticipate the possibility that this will not always be so, and that there may well be cases (such as abortion) where Parliament will assert its political and legal authority over the courts. In this way the sovereignty of Parliament will be preserved both in principle and in practice, even though it is unquestionably the case that Parliament's position would have been stronger politically and consitutionally if incorporation of the ECHR had stopped short of giving the courts the power to challenge primary legislation. But by ensuring that legal sovereignty continues to rest with Parliament, the government has retrieved the first constitutional principle of democratic socialism, and overcome one of the principal concerns about incorporation of the Convention. It now remains to be seen for how long a wounded Act of Parliament will be able to survive a declaration of incompatibility to which the government and Parliament refuse to respond.

125 *ibid*, col 1301.

© The Modern Law Review Limited 1999

Law and Landscape: The Legal Construction and Protection of Hedgerows

Jane Holder*

Law has left enduring marks on the landscape and continues to shape the physical environment. The recognition of this has been slow in coming, law being relatively silent on landscape and typically more concerned with controlling behaviour and relationships than structuring and ordering the form of land. However, as links between the critical traditions in law and geography are made,[1] so the relationship between law and landscape becomes a little clearer. In particular, the legal construction of landscape reveals law's physicality in a similar manner to that in which its materiality has been exposed.

'Landscape' has many meanings.[2] At least two might be mentioned here in order to provide a schema by which law's relationship with landscape may be explored. The first refers to the surface features of an area, including both natural elements (eg landform, water, soil, flora and fauna) and artificial objects (eg settlements, roads, dams, canals) shaped by natural processes and human activities. The second meaning is that landscape is the group of visible features of an area from an observer's viewpoint, defined by forms, lines, colours and textures, all of which provide a framework for the interpretation of nature and culture. The latter recognises that landscape stimulates the visual senses and may alter according to individuals' differing perspectives, histories and experiences. In this sense, landscape is inseparable from human perception; it is 'the work of the mind'.[3] Appreciation of the indivisibility of landscape and people was at the core of the original meaning of the Dutch word 'landschap', meaning a unit or jurisdiction of human occupation which entered the English language as 'landskip' at the end of the sixteenth century.[4] This prevails in the idea of cultural landscapes – landscape as a product of interaction between people and nature. Indeed, very few landscapes have not been touched by people: some have been artifically created, others transformed and shaped by human practices and land uses.[5] Thinking more imaginatively, rather than landscapes representing or being shaped by culture, the set of values, beliefs and ideas held by a social group

* Faculty of Laws, University College London

I would like to thank Donald McGillivray, Colin Scott, Sue Elworthy and Clare Tonkins for their help and comments.

1 See N. Blomley, *Law, Space and the Geographies of Power* (New York: Guilford, 1994) and D Cooper, *Governing Out of Order: Space, Law and the Politics of Belonging* (London: Rivers Oram, 1998).

2 For a discussion of those discussed here and other meanings, see A. Kiss and D Shelton, *Manual of European Environmental Law* (Cambridge: Cambridge University Press, 1997) ch 11

3 S. Schama, *Landscape and Memory* (London: Fontana, 1996) 7; see also the account of the Greek technique of memory by impressing places and images on the mind in F Yates, *The Art of Memory* (London: Routledge and Kegan Paul, 1966).

4 Schama, n 3 above, 10.

5 G. Bennett, *Cultural Landscapes. the Cultural Challenge in a Changing Europe* (London: Institute for European Environmental Policy, 1996).

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

at a particular time and place ('mental products'), landscape may be thought *to be* culture. This understanding is a development of the second meaning of landscape: because observation or perception makes the difference between raw matter and landscape, differentiating between a natural landscape and a cultural landscape makes little sense, unless, of course, non-human values and interests are at issue. Landscape therefore has considerable cultural power. Accordingly, landscape has been used to invoke the major concerns of modernity – nation, freedom, liberty, and state – in ways which may be dangerous when abused.[6] For example, following the deportation and killing of Jews in southern Lithuanian woodland villages the landscape was to be altered so that it became something unmistakably German, organised by the Reichskommissar for the Affirmation of German Culture. This was to be coupled with the expansion of forest in Lithuania, stocked with Teutonic species, some of which such as the Bison, were reclassified for the purpose. This amounted to a plan to carve German identity from a quite different topography.[7]

This article uses the case of the hedgerow to show that law has played an important part in the construction of the rural landscape, but less so in its protection. This is partly because law appears to conceive of landscape only in the first sense of surface features of an area. This means that a sense of landscape as perspective, and cultural signifier, is absent in legal regimes for its protection. The main concern of the article, law's shaping of the physical environment, specifically the rural landscape, and its evocation of landscape in a set of legal rules[8] feeds into current debates on the state of the environment and its protection. First, using the hedgerow to trace the legal construction and protection of landscape reveals law's part in defining what is meant by 'nature'. I examine the way in which law considers hedgerows to be little more than the sum of their parts – trees, fauna and flora – and 'nature' as constitutive of these. The cultural significance of the hedgerow, as an essential part of an idealised rural picture, and as representative of a particular locality is not registered. The main argument is that by failing to take account of these aspects of the social and cultural importance of the hedgerow, law only narrowly views the meanings of 'natural' attributed to them, and thereby ignores that nature may be created by culture, the categories of the natural always being open.[9]

The second issue raised by this article is the status of hedgerows as commons. Paradoxically, although many hedgerows were planted to give effect to enclosure by excluding peasants from common lands and demarcating the boundaries of property, they are now considered to be 'commons' themselves, part of the shared natural heritage.[10] There is therefore some continuity between the issues and problems which arose from their planting and those which arise from their loss in modern times: the meaning and nature of threats to the commons; the exclusion of groups from decisions about the future of land;[11] and the capability of law, typically concerned with the protection of individual property rights, to protect 'commons'.

---

6 See I. Coates, 'A Cuckoo in the Nest: the National Front and Green Ideology', in J. Holder *et al* (eds) *Perspectives on the Environment* (Aldershot: Avebury, 1993).

7 Schama, n 3 above, 70–71

8 In contrast to a concern with the structuring of relationships within the countryside, which has been explored more fully in other work, for example D. Harvey, *Justice, Nature and the Geography of Difference* (London: Blackwell, 1996) and D. Cooper, *Governing Out of Order*, n 1 above.

9 See N. Evernden, *The Social Creation of Nature* (Baltimore: Johns Hopkins University Press, 1993).

10 Countryside Commission, *Views From the Past: Historic Landscape Character in the English Countryside* (Cheltenham: Countryside Commission, 1994).

11 A similar example is the sale of land by the Forestry Commission, 'Ministers urged to halt forest sales', *The Guardian*, 4 February 1998.

© The Modern Law Review Limited 1999

I discuss law's shaping of the physical environment and the issues outlined above in three parts. First, I locate the hedgerow in the landscape by considering its historical, ecological and cultural importance. Second, I explore the legal construction of landscape, using the example of the mass planting of hedgerows which was a consequence of the Enclosure Acts. In the third section I examine legal mechanisms for the protection of landscape. Again focusing on hedgerows, I review the main, but belated, legislative response to their loss, the Hedgerows Regulations 1997.[12] This secures protection only for those hedgerows deemed to be 'important' according to quantifiable criteria (How many species? How many features? How old?) rather than by their cultural or local importance. I consider that this approach typifies the narrowness of legal definitions and conceptions of landscape, particularly when compared with non-legal meanings and understandings, and those which emphasise the relationships between people, nature, and landscape. The general issue is that the Regulations represent a shift towards calculability in law. Even before the Regulations were made, however, the recognition of the conservation value of the hedgerow meant that laws, some ancient, were employed by individuals. I compare this piecemeal, but creative, use of law with the protection accorded by the 1997 Regulations. By way of conclusion, I discuss the asymmetry of law's role in constructing and protecting the landscape, particularly the incongruity of conserving 'commons' such as hedgerows via legal mechanisms designed to protect private property.

Placing the hedgerow in landscape

As ancient forests began to be cleared for agriculture in Prehistoric Britain, hedges were left as the skeletons of these woods, their edges creating field boundaries. In late Saxon and Norman times (c 440–1066) more woodland was removed to clear space for open field strip farming, leaving hedges as barriers to control the movement of livestock.[13] Beyond the open fields was common land subject to shared rights by villagers to gather wood and berries, to mow grass for hay to feed their stock and for grazing. Many of these open fields and commons were enclosed by hedges which followed the lie of the land and trackways. Some hedgerows also developed naturally at the edges of fields when a fence, ditch or bank near a source of tree seed was neglected for a few years. Hedgerows were the normal furniture of the rural landscape. They provide evidence of centuries-old farming systems, particularly the mixed farming of lowland Britain, and a link with past rural traditions and crafts.[14]

Hedgerows have traditionally supported a dense and diverse fauna and flora, acting as a reservoir of wildlife. Elder, blackthorn, privet, crab apple, bullace, and sloe, oak, ash, willow and field maple are some of the plants and trees which adapted and survived in the hedgerows which grew up as woodland was cleared for human settlements. For birds (eg skylark, hedge sparrow, chaffinch, song thrush, yellow hammer, and wren), animals (eg common rat, fox, stoat, badger, weasel, hedgehog, mole and shrew) and invertebrates (eg butterflies, earthworms,

12 SI 1997 No 1160.

13 See S. Elworthy, *Farming in Britain and its Relationship with the Environment* (unpub LLM dissertation, University of Warwick, 1990), although the extent of the clearances by the Angles and Saxons has been exaggerated. O. Rackham, *Trees and Woodlands in the British Landscape: the Complete History of Britain's Trees, Woods and Hedgerows* (London: Routledge, 1996) 56.

14 W.H. Dowdeswell, *Hedgerows and Verges* (London: Allen and Unwin, 1987).

© The Modern Law Review Limited 1999

centipedes, beetles and aphids) hedgerows similarly provided an extension of the ecological conditions of woodland. By providing corridors of habitat linking woodland that would otherwise be isolated, hedgerows allowed them to extend their range and so avoid the pressures of competition.

The ecology of the hedgerow is closely related to its cultural importance. For example there are many cultural associations with its produce – the berries (barberries, blackberries, elderberries, raspberries and strawberries) and weeds, fungi, herbs, greens, flowers and nuts, and also the identification with particular local styles of hedgerow (beech hedges in Exmoor, gorse hedges in west Wales). In this respect the cultural landscape of fields and hedgerows might give meaning to ideas of the 'rootedness' of a community,[15] not necessarily in a geographical sense but in terms of people's aesthetic and material experiences of landscape as a sense of place,[16] for example, of planting or foraging a hedgerow. In turn, cultural influences may be seen to determine the ecology of hedgerows. The 'management' of hedgerows by those living and working in the landscape is crucial to the survival of the hedgerow and its inhabitants, particularly given the rate at which 'traditional', or rural ways of life have adapted or succumbed to modern, industrialised society.[17] This suggests that the issue of conservation should be logically extended to include consideration of human social and cultural diversity also.[18]

The hedgerow is commonly portrayed as a quintessential, natural, feature of the English landscape in literature, poetry, paintings and maps[19] even though there is nothing distinctly English or indeed British about hedgerows.[20] The idea of Elysian rural England has been taken up by English socialists and liberals, but mostly conservatives. For example, towards the end of the nineteenth century a group of Tory poets created in their work a picture of unchanging England to set against rising social unrest and foreign threats.[21] Amongst this group, Poet Laureate Alfred Austin idealised rustic, old England in his 'Return to England' which pictured hedgerows, cottages and village spires.[22] By the First World War, nostalgic and utopian visions of the country encouraged patriotism. An anthology of writing, *England, the Old Country*, was distributed to soldiers evoking the England of Alfred Austin and Rupert Brooke. 'O England', E.V. Lucas' poem in the collection sentimentalised 'country of my heart's desire, land of the hedgerow and the village spire'. Even in the 1990s hedges and fields still form an essential part of an idealised, rural myth or 'national essence' which serves to uphold certain social hierarchies and practices, most obviously hunting.[23] This is combined with the

15 See, for example, R. Cotterrell, *Law's Community: Law's Theory in Sociological Perspective* (Oxford: Clarendon, 1995) and B De Sousa Santos, 'Law: A Map of Misreading: Towards a Post Modern Conception of Law' (1987) 14 *Journal of Law and Society*, 279.

16 See H. Hannum (ed), *People, Land and Community* (New Haven/London: Yale University Press, 1997) and P. Hepburn, 'Landscape and the Metaphysical Imagination' (1996) 5 *Environmental Values* 3.

17 See for example, Bennet, n 5 above

18 See J.F. Thorne, and C.S. Huang, 'Toward a Landscape Ecology Aesthetic: Methodologies for Designers and Planners' (1991) 21 *Landscape and Urban Planning* 61, 61–63.

19 Schama, n 3 above, and D. Lowenthal, 'British National Identity and the English Landscape', (1991) 2 *Rural History* 205.

20 For example, Western Normandy is *bocage* country, typified by intricate and well established hedges, and there are hedged areas in western Italy, the Austrian Alps and Greece: O. Rackham, *The History of the Countryside* (London: Dent, 1986) 181.

21 M. Weiner, *English Culture and the Decline of the Industrial Spirit 1850–1980* (Cambridge: Cambridge University Press, 1981) 55–56.

22 *Lyrical Poems* (London: 1891).

23 D. Cooper, 'Hunting, Land and Judicial Intervention', in D. McGillivray and J. Holder (eds) *Locality and Identity: Environmental Issues in Law and Society* (Aldershot. Dartmouth, 1999).

© The Modern Law Review Limited 1999

ideals of some environmentalists who view the health of the hedge as a litmus for modern farming practices. A further aspect of the modern cultural significance of the hedgerow is the activities of an 'Anti-Leylandii' guard who wish to use law to protect against the rampant growing of neighbours' cypresses in place of the traditional hedge.[24]

Throughout history the cultural associations of the hedgerow with 'nature' have proved to be complex. John Stuart Mill bemoaned the widespread uprooting of hedgerows in the name of agriculture, which he saw as part of farming's master plan to tame nature: 'with every foot of land brought into cultivation ... every hedgerow or superfluous tree rooted out, and scarcely a place left where a wild shrub or flower could grow without being eradicated as a weed in the name of agriculture'.[25] Such a portrayal belies the artificiality and ambiguity of hedgerows. Although some hedgerows are naturally occurring, most were originally planted to divide fields and mark the boundaries of property. Confusingly, even these tended to be rooted on ancient wood boundaries. Hedgerows have also shown a capacity for attracting flora other than those planted, so that they take on different ecological characters, a propensity now greatly limited. When originally planted in great numbers in the eighteenth century, hedgerows were seen as monotonous and unnatural, an indication that agriculture, like the towns, was moving away from nature.[26] William Gilpin expressed this fear: 'wherever man appears with his tools, deformity follows his steps. His spade and his plough, his hedge and his furrow, make shocking encroachments on the simplicity and elegance of landscape'.[27] Representing both the domination and presence of nature, the hedgerow exposes the transience of 'nature'; that it is shaped by culture and time.

Legal construction of landscape: planting hedgerows

Changes in the rural landscape may be explained primarily as a result of social and economic forces, and particularly advances in agricultural technique. But law has also been used both consciously and unconsciously to shape the landscape. Open fields, with their ridges and furrows and bounded by hedges, were established and regulated by customary laws administered by local manorial courts. Hedgerows have long been planted to mark the boundaries of property, and hence were often mentioned in charters for the sale or gift of land.[28] However, the mass planting of hedgerows occurred to give effect to the enclosure movement by demarcating the enclosed land (the old open fields) and fencing off the 'waste' or common lands, leading to the abolition of common land rights and husbandry practices, and large tenanted farms.[29] This process began roughly in 1450 and reached a peak between

24 M. Jones, 'Growing Pains: Diary of an Anti-Hedge Activist' *The Guardian*, 13 May 1998. The author has set up 'Hedgeline'.

25 J.S. Mill, *Principles of Political Economy* (London: Penguin, 1848, reprinted in 1970) 58

26 D.A. Evans, A *History of Nature Conservation in Britain* (London: Routledge, 1992). 27.

27 W. Gilpin, *Remarks on Forest Scenery and Other Woodland Views (Related Chiefly to Picturesque Beauty), in Three Books* (London: R. Blamire, 3rd ed, 1808, reprinted by Richmond Publishing Co Ltd. Surrey, 1973), 38.

28 For example, *Boundary of Hurstbourne Priors* (Hants) AD 910: 'Start from Twyford along the road to Bracken Ridge. then to the road that shoots over the ditch; then along the road from where it adjoins the wood...by the little hedge along the spinney...along the hedge to the old maple tree', quoted in Rackham, n 13 above, 44

29 Rackham, n 13 above, 196, estimates that 200,000 miles of hedge was planted between 1750 and 1850; see also Dowdeswell, n 14 above, 15.

© The Modern Law Review Limited 1999

1750 and 1850.[30] There were many reasons for enclosure, mainly of an economic nature, including the need to secure more commercially oriented farms so as to provide for an increasing urban population. It is possible also that enclosure was a response to environmental degradation caused by the open field system which exhausted the arable land, and led to depopulation in rural areas.[31]

The role of law in the process of enclosure was to ratify and legitimate agreements between proprietors to restore their land in its entirety.[32] One method was to obtain a Chancery Decree, a necessary fiction of which was that a dispute between proprietors existed. When the Chancery Court ruled in favour of one or other of the parties, the validity of the enclosure was firmly established. Among these collusive disputes were genuine disagreements, for example where an attempt was made to enforce enclosure on those who had reneged on an agreement or had never made one. The Chancery Decree was the significant legal form of early enclosure which tended to be instigated by the act of an individual and was a piecemeal and disorderly means of appropriating common field lands. This form was succeeded as a means of ratifying enclosures by Parliamentary Acts which explicitly allowed enclosures to which some of the proprietors had not given their consent.[33] This permitted the more orderly and far-reaching process of enclosure of land. Until the General Enclosure Act 1836, each case of enclosure was considered separately by Parliament and generally required a local Act. Thereafter, enclosure was permitted without a local Act if two-thirds of the proprietors of the land gave their consent. Both local and Parliamentary Enclosure Acts were therefore used to confirm existing arrangements and extend landlords' control over the land and its inhabitants. This was generally arranged among large landowners and its acceptability determined in Parliament by other members of that class. In this way, law operated as a legitimating device for decisions already taken about the restructuring of land and property relations.

The different legal forms adopted to achieve enclosure had a direct bearing on the physical layout of the land and the landscape. The early, piecemeal ('do it yourself') type of enclosure in which individual landowners enclosed their own strip of land from the common lands by way of a Chancery Decree led to the creation of small, irregular fields with hedges following natural boundaries. In contrast, the landscape of the Parliamentary enclosures, giving effect to enclosure by general, or majority consent, was more planned and had a mass produced quality, with straight hedges and rectangular fields. The Enclosure Acts therefore provide the main example of how law has borne upon the landscape and shaped it. The pastoral scene of fields divided by hedges and punctuated with woods and thickets is a legally engineered landscape, shaped by the creation, then upholding, of private property interests and the legal denial of common land rights.

Legal protection of landscape: conserving the hedgerow

Following the Enclosure Acts, the rural landscape was relatively stable: the majority of hedges which existed in 1870 were still there in 1950.[34] All then

30 J.V Beckett, *The Agricultural Revolution* (Oxford: Blackwell, 1990) 35

31 See H. Bradley, *The Enclosures in England: An Economic Reconstruction* (New York: AMS, 1968)

32 J A Yelling, *Common Fields and Enclosure in England 1450–1850* (London: Macmillan, 1977) 8 *et seq*

33 H.S. Homer, *An Essay on the Nature and Method of Ascertaining the Specifik Shares of Proprietors Upon the Inclosure of Common Fields* (Oxford: 1766) quoted in Yelling, *ibid* 8.

34 Rackham, n 20 above, 191

© The Modern Law Review Limited 1999

changes. The economics of agriculture, responsible in the main for the enclosure movement centuries earlier, led to a greater emphasis on arable farming rather than rotational farming which had relied on hedgerows to prevent livestock straying onto crops. The focus on high production levels to maximise subsidies resulted in increased use of fertilisers and pesticides and large farm machinery such as combine harvesters with wide turning circles. Small fields that had previously housed livestock were no longer economic and so it became common practice for the size of these fields to be increased by grubbing out hedges. Such technological farming was encouraged by the Ministry of Agriculture, Fisheries and Food which, until 1992, gave grant aid to farmers to remove their hedges. This practice reached a peak in the mid 1960s, with the loss of about 10,000 miles of hedgerows per year.[35] The rate of loss has since slowed[36] but hedgerow removal continues, causing loss of species and making the resultant large fields vulnerable to wind erosion of the fertile top soil as well as water erosion.

Although the removal of hedgerows has been brutal, the greater threat to the hedgerow is from dieback, caused by local environmental factors arising from modern agricultural practices. For example, ploughing close to trees damages root systems, compaction of soil by farm vehicles results in a breakdown in soil structure and reduced root function and heavy application of fertilisers leads to growth abnormalities.[37] The declining standard of management of hedgerows also bears some responsibility for dieback. Mechanised flails bring a 'short back and sides' look to the hedgerow, which fails to protect saplings. This compares with a time when hedges were cut manually with billhooks into intricate patterns that provided for new growth and allowed for the tailoring of hedgerows to the local landscape so that Welsh, Midlands, and West Country hedgerows all differed.

In response to these influences, law has traditionally upheld, not conservation of the landscape, but rather good husbandry and estate management. The Agriculture Act 1947 which laid the basis for the post-War agricultural arrangements required that the necessary work of maintenance and repair of hedgerows was to be carried out, for example hedges were to be cut and laid rather than left to become overgrown.[38] The 1947 Act was the product of the Report of the Committee on Land Utilisation in Rural Areas (the Scott Committee) (1943) which proposed that farmers should, by virtue of their ordinary agricultural practices, continue to be custodians of the landscape, holding it in trust for the whole nation:

> Farmers and foresters are unconsciously the nation's landscape gardeners...even were there no economic, social or strategic reasons for the maintenance of agriculture, the cheapest way, indeed the only way, of preserving the countryside in anything like its traditional aspect would still be to farm it.[39]

In the event, the 1947 Act was too weak to control the direction of agriculture.

---

35 Countryside Commission, *New Agricultural Landscapes* (Cheltenham: Countryside Commission, 1977) para 9. A report by the Institute of Terrestrial Ecology on hedgerow changes in Great Britain between 1984 and 1990 revealed that the total hedgerow length in England had fallen in the period by over 20 per cent and in Wales by 25 per cent: C.J. Barr *et al, Changes in Hedgerows in Britain 1984–1990* Department of the Environment/Institute of Terrestrial Ecology (London: HMSO, 1991)

36 The Institute of Terrestrial Ecology estimate that between 1990 and 1993 600 km of hedgerow was lost each year. The rate of gains from new planting increased in this period from 1,900 km to 4,400 km per year: C Barr, M. Gillespie, D Howard, *Hedgerow Survey 1993* (Grange over Sands: Institute of Terrestrial Ecology, 1994)

37 Forestry Commission, *Ash Dieback*, Bulletin 93 (London: HMSO, 1991)

38 Agriculture Act 1947, s 11(2)(d).

39 Committee on Land Utilisation in Rural Areas (the Scott Committee) *Report on Land Use in Rural Areas*, Cmd 6378 (London: HMSO, 1943) para 160.

© The Modern Law Review Limited 1999

The Town and Country Planning Act 1947 was similarly a response to the approach adopted by the majority of the Scott Committee. As a consequence, agricultural and forestry (rural) land uses, including those affecting hedgerows, did not come within the ambit of the British development control system,[40] this urban bias suggesting that the Act's title was something of a misnomer. The removal of hedges in particular illustrated the limitations of planning law and led to arguments that the planning system should protect hedgerows in the same way as it does certain trees.[41] Instead, protection of hedgerows has (until recently) been advanced by aid for their management, utilising economic measures such as the Hedgerow Incentive Scheme[42] (which replaced the grant aid given to farmers for removing hedges) with its provision for the negotiation of management agreements between farmers and the Countryside Commission for planting, restoring and maintaining hedges.

The part to be played by law in protecting landscape has been slowly recognised, as illustrated by the formulation of statutory duties which, variously, require public bodies to have regard to the beauty and amenity of the landscape.[43] However, such duties are generally not combined with compulsory powers to support landscape protection, grant aided management agreements with owners remaining the main means of achieving this aim.

More specifically, the law lagged behind the reality of the loss of hedgerows. As farming came to be seen as increasingly problematic in conservation terms several private members' bills were brought to introduce specific protection regimes for hedgerows. These failed to be enacted, as did a government proposal that farmers be paid to preserve rather than remove hedgerows.[44] Taking the opportunity offered by the passage of the Environment Act 1995, the government finally gave itself powers to make regulations to protect hedgerows,[45] the result of which is the Hedgerows Regulations 1997.[46] In so doing the government was influenced by commitments to secure biodiversity in international law and EC law[47] and to address concerns about the rural landscape.[48] The Regulations were approved by Parliament just before the general election. Several weeks into office, the Environment Minister, Michael

40 Town and Country Planning Act 1990, s 55(1)(e) excludes forestry and agricultural land uses from the scope of development control.

41 For example, under the Town and Country Planning Act 1990, s 197, the local planning authority must ensure that in granting planning permission for development adequate provision is made by the imposition of conditions for the preservation or planting of trees. Under s 198 of the 1990 Act the local planning authority may make a tree preservation order which may prohibit the cutting down, topping, lopping, uprooting or destruction of trees.

42 Introduced in 1992, and now part of the Countryside Stewardship Scheme, administered by the Ministry of Agriculture, Fisheries and Food.

43 Wildlife and Countryside (Amendment) Act 1985, s 4, required that the Forestry Commission take account of such features by inserting s 1(3A) into the Forestry Act 1967. Similar aims were introduced into planning legislation which required that development plans include policies on 'the conservation of natural beauty and amenity of the land': Town and Country Planning Act 1990, ss 12(3A), 31(3), 36(3). The Environment Act 1995, s 61 substitutes a new s 5 in the National Parks and Access to Countryside Act 1949 which requires all public authorities to have regard to 'conserving and enhancing the natural beauty, wildlife and cultural heritage of the areas'.

44 Planning and Compensation Bill, HL Deb vol 54 Written Answers cols 85–87.

45 Environment Act 1995, s 97

46 SI 1997 No 1160.

47 See, eg Department of the Environment, *Biodiversity. the United Kingdom Steering Group Report Vol 2: Action Plans* (HMSO: London, 1995) which includes an action plan for the protection of certain ancient or species rich hedgerows.

48 HM Government, *Rural England*, Cm 3016 (London: HMSO, 1995)

© The Modern Law Review Limited 1999

Meacher, announced a review of the Regulations, indicating that some doubts existed as to their adequacy and efficacy.[49]

The 1997 Regulations prohibit the removal of hedgerows unless a procedure is followed, thus finally separating arrangements for the protection of hedgerows from aid for their management. Owners must give notice to the local planning authority that they wish to remove a hedgerow.[50] Should the planning authority decide, according to criteria set by the Minister, that the threatened hedgerow is 'important', it may issue a hedgerow retention notice stating that the works are not to be carried out. The system is reinforced by criminal sanctions and the requirement that hedgerows removed in contravention of the Regulations must be restored. The Regulations represent a new direction for conservation law. The legal regime – of notification, license, and sanction – is the familiar 'command and control' model pursued for development control, but is here invoked for the purposes of protecting hedgerows as part of the rural landscape. This legal form usurps the tradition of governing the countryside via 'management' by farmers, backed up by voluntary systems of financial incentives, grants and subsidies.[51]

The key to the operation of the Regulations is the defining of 'important' hedgerows by Ministers. These are currently considered in the Regulations to be hedgerows which have existed for 30 years or more and which satisfy at least one of two sets of criteria.[52] The first set relates to archaeological and historical matters, for example that the hedgerow marks the boundary of at least one historic parish or township existing before 1850. The second set relates to wildlife and landscape considerations and includes lists of woodland species (bluebell, oxlip, toothwort etc), woody species (rowan, alder, wild cotoneaster etc), endangered or vulnerable animals and birds species,[53] and various hedgerow features. The criteria are complex, precise and technical and provide an extreme example of the specificity of regulatory rules.[54]

The general approach of the 1997 Regulations of very closely defining the type of hedgerow deserving legal protection is flawed in several respects. First, only the proposed *removal* of a hedgerow triggers legal protection. This includes within its scope acts that result in the destruction of the hedgerow, for example a deliberate act carried out on nearby land. Slow dieback, caused by less direct and obvious acts such as the over-application of chemicals, which poses the main threat to hedgerows, is not captured by the Regulations. Second, the central concept of the Regulations, the protection of 'important' hedgerows, might be questioned on the ground that this secures protection for the 'fittest' hedgerows in terms of species inhabiting them and the features that they boast. However, once hedges have deteriorated to the point that they are of very poor quality, there are no farming reasons to keep them. And a grubbed out hedgerow stands no chance of becoming an 'important' hedgerow over time. The Department of the Environment, Transport and the Regions' *Review of the 1997 Hedgerow Regulations* also points

49 Department of Environment, Transport and the Regions, *Review of the Hedgerows Regulations 1997* (London: DETR, 1998).

50 Reg 5.

51 See S. Bell, *Environmental Law* (London: Blackstone Press, 3rd ed, 1997) ch 16

52 Reg 4. The criteria are set out in Part II of Sched 1 to the Regulations

53 These include birds and animals protected by the Wildlife and Countryside Act 1981, or categorised as a declining breeder or endangered, extinct, rare or vulnerable in various lists drawn up by the Royal Society for the Protection of Birds or the Nature Conservancy Council.

54 See R. Baldwin, *Rules and Government* (Oxford: Oxford University Press, 1995).

© The Modern Law Review Limited 1999

out that only countryside hedges are protected and recommends that protection be extended to hedges in urban areas.[55]

The most significant flaw is that the criteria by which hedgerows are judged to be important are very narrow. This is best illustrated by the factors to be considered when judging the wildlife and landscape importance of a hedgerow, particularly that the criteria are couched in quantifiable terms. A hedgerow is defined as important if it includes at least seven woody species listed in the Regulations. This need only be six species where at least three specified features exist. These features are listed as a ditch along at least one half of the length of the hedgerow, a standard tree for each 50m of hedge, or certain connections with other natural features, particularly other hedgerows (hedgerows forming part of a continuous network are recognised as having special value).[56] In the case of the last: 'a connection with another hedgerow scores one point and a connection with a pond or a woodland in which the majority of trees are broad leaved trees scores 2 points'.[57] Connections 'scoring 4 points or more' constitute a 'feature' of the hedgerow.[58] The operation of these criteria presupposes a scientific assessment of the number and the type of species present in a measured stretch of hedgerow. The underlying premise is that 'importance' *can* be quantified in this manner; the criteria do not act as a proxy for more subjective considerations. This corresponds to an approach to conservation by which values are attributed to natural resources or 'capital' by quantifiable terms, for example through individuals' willingness to pay money to protect or enhance nature.[59] The focus of this approach is on 'features' or 'objects' requiring protection rather than the recognition of less tangible attributes such as typicalness, distinctiveness, setting and context.[60] The adoption of quantitative criteria in the 1997 Regulations may be seen as forming part of the more general incorporation of economic valuing techniques in environmental law.[61] The main issue is that the Regulations focus on calculability, so that the non-calculable diminishes in value.[62]

This approach of calculating the conservation value of hedgerows is combined in the Regulations with collecting and counting rare species, a preoccupation of early conservationists[63] and criticised in the context of landscape protection, for example by the Countryside Commission:

> Most people in Britain now live in towns and cities and, although few may interpret the visual quality of the landscape in specific terms, we can say with confidence that millions derive pleasure from it ... there is no way in which we can quantify or evaluate this pleasure. It is sensual and individual. It has been the inspiration for much of our music, painting, poetry and literature. Our countryside is a part of our heritage, and we hold it in trust.[64]

---

55 Department of the Environment, Transport and the Regions, n 49 above, para 6.10.

56 Para 7(1)(a) and (b) of Schedule I, Part II, to be read in conjunction with para 7(4)(h).

57 Para 7(5) of Sched 1, Part II

58 Para 7(4)(h) of Sched 1, Part II. The working of these criteria means that the percentage of hedgerows currently protected is about 20 per cent in England and Wales, though the impact of the criteria varies in different parts of the country (Department of Environment, Transport and the Regions, n 49 above, para 1.5).

59 As advanced by D. Pearce, *Blueprint for a Green Economy* (London: Earthscan, 1989).

60 See C. Harrison, J. Burgess and J. Clark, 'Capturing Values for Nature: Economic and Cultural Perspectives' in McGillivray and Holder, n 23 above.

61 For example, the introduction of cost-benefit analysis as a legal requirement by the Environment Act 1995, ss 4 and 39, and the concept of Best Available Techniques Not Entailing Excessive Cost (BATNEEC) in the Environmental Protection Act 1990, s 7(7).

62 On calculability in law, see M. Power, *The Audit Society* (Oxford: Oxford University Press, 1997) and P. Miller and M. Power, 'Accounting, Law, and Economic Calculation' in M Bromwich and A. Hopwood, *Accounting and Law* (Hemel Hempstead: Prentice Hall, 1992).

63 Evans, n 26 above, 53–57.

64 Countryside Commission, n 35 above, para 21–1

© The Modern Law Review Limited 1999

The narrowness of the criteria to determine 'important' hedgerows can best be seen by the failure of the Regulations to acknowledge the cultural significance of hedgerows. For example, no mention is made of the significance of their differing styles in various regions, and hence their representativeness of locality.[65] This shortcoming is attributed in the review of the Regulations to the setting of criteria to determine 'important' hedgerows at the national level and it is therefore recommended that local authorities determine which hedgerows in their areas should be protected.[66] The drafters of the Regulations perhaps considered that the cultural significance of hedgerows stems from their historical or ecological importance. This is the case in part. However, the absence of discrete cultural criteria means that no regard may be had to local feeling about a hedgerow by means, for example, of public consultation.[67] Instead, the Regulations advance conducting a uniform audit of the species and features of the hedgerow and surrounding area according to centrally set criteria. Such weak protection for culturally significant, particularly locally representative, hedgerows which might not fulfil the existing ecological, wildlife and historical criteria, underlines a need for law to be more locally applicable and adaptable. From this perspective, the management agreement, allowing for some degree of negotiation between parties and variability according to local conditions, might be a more appropriate mechanism for protecting landscape.

The current treatment of hedgerows in the 1997 Regulations suggests that law conceives of hedgerows in terms of the first meaning of landscape outlined in the introduction to this article – the surface features of an area, the easily definable and quantifiable characteristics. In contrast, the second meaning of landscape, the *observer's* intuitive perspective of features of an area – the landscape as culturally defined – is absent from the legal regime for protecting hedgerows. The general picture of law derived from the Regulations is of a set of complex and technical rules which are simultaneously simplistic and superficial, the preoccupation being with counting (of species, length, and features) rather than the meanings to be attributed to hedgerows, and deriving from them. Underlying this is scant understanding of the place of people in the protection of landscape, the connections between hedgerows counting for more than the connections between people, nature and landscape. The hallmarks of modern law are all present in the 1997 Regulations: universalism, disembededness (from the social and cultural milieu of a particular community), and instrumentalism.[68]

Law has proved itself not entirely incapable of taking account of cultural, or intangible, values.[69] Nevertheless, the main advances in the incorporation of cultural values in decision making processes are being made outside law, the general aim of which is to develop more deliberative and inclusionary methods of

65 Interestingly, the proposed amendment to the Planning and Compensation Bill for the legal protection of hedgerows did include a reference to 'having regard to local custom and practice'.

66 Department of Environment, Transport and the Regions, n 49 above, paras 1.5–1.7.

67 The provisions on consultation in the Regulations are very narrow: Reg 5(3) requires that the local planning authority consult the relevant parish council on hedgerow notices, but this is to determine the importance of certain hedgerows in terms of the criteria listed.

68 T. Banuria and F. Marglin, *Who Shall Save the Forests? Knowledge, Power and Environmental Destruction* (London: Zed Books, 1993) 9–13.

69 For example, in international law the Convention for the Protection of the World Cultural and Natural Heritage aspires to safeguard that 'unique and irreplaceable property which is part of the cultural heritage'. However, in the case of *enforceable* procedural rules, the status of interests of a cultural character rank lower than interests of a more tangible character, such as the protection of the individual and the material value of property, see J. Riley, '*Locus Standi* and Cultural Interests' (1997) *Journal of Planning and Environmental Law*, 20.

 © The Modern Law Review Limited 1999

valuing nature, in particular involving the public. These developments might be located within 'common good' approaches to valuing nature which implies embracing ethical and moral concerns as well as popular values for particular places, features or their attributes.[70] For example, the Countryside Commission has developed a new map of the English landscape with the aim of defining discrete regional character areas. Alongside such attributes as geology and soils, cultural associations were also drawn upon by exploring the public's perception of the local landscape. This new map of landscape, viewed from a regional perspective, will be used to identify local priorities and develop a series of regional landscape conservation strategies.[71] This is not to say that the incorporation of cultural values (those held by a community or locality) *in law* for the protection of landscape would not be unproblematic. For example, since the cultural appreciation of hedgerows has changed considerably over time there is some danger in protecting hedgerows as one detail of the 'romance of the countryside' at the cost of failing to recognise the potential significance of new forms in landscape such as the motorway verge. Defining 'locality' or 'community' in law is also difficult since these concepts are also subject to change. More importantly, giving legal expression to local and community cultural values might contribute to the safeguarding of a rural culture which may be exclusionary.[72]

The limitations of law in accommodating cultural values may also be attributed to the possibly undesirable goal of trying to take account of the incalculable in law. This view is held by Harte who considers that no system of law can by itself create an attractive landscape or control land so as to satisfy everyone in society. He argues that the more legislators try to trap elusive and subjective qualities such as beauty or amenity in a framework of rules, the less clear the law becomes and the more initiative is stifled.[73] On the contrary, 'elusive' might be taken to mean cultural values – knowledge and aesthetic experience – which may, in jurisprudential terms, be regarded as 'basic goods'.[74] From this perspective, it is the detailed knowledge and 'management' practices of people which preserves and sustains a landscape, not the scientific recording of its calculable features. Conservation law and policy which fails to take account of this is inevitably half-hearted. It might even be questioned whether, by writing culture out of the story of landscape protection, law is producing an 'anti-landscape', so closely tied are cultural values and landscape.[75]

More practically, the Hedgerows Regulations 1997 demonstrate a regulatory paradox: the leisurely speed at which the government proceeded towards legislation encouraged landowners to hasten destruction of hedgerows prior to the legislation taking effect.[76] The tenor of the accompanying 'good practice' guide[77] for local planning authorities charged with administering the system is inauspicious. This emphasises that the retention of important countryside hedgerows is most likely to be met by adopting a common sense approach, based

70 See Harrison, Burgess and Clark, n 60 above.
71 Countryside Commission, *The New Map of England: A Celebration of the South Western Landscape* (Cheltenham: Countryside Commission, 1994).
72 For example, see Cooper, n 23 above.
73 J.D.C. Harte, *Landscape, Land Use and Law* (London/New York: Spon, 1985).
74 J. Finnis, *Natural Law and Natural Rights* (Oxford: Clarendon Press, 1980).
75 Schama, n 3 above, 61.
76 Since legislation was promised in 1990, at least 60,000 miles have been grubbed up or have degenerated through neglect, Leader article, *The Guardian*, 3 January 1997.
77 Department of the Environment/Ministry of Agriculture, Fisheries and Food, *The Hedgerows Regulations 1997: A Guide to the Law and Good Practice* (London: HMSO, 1997).

© The Modern Law Review Limited 1999

on cooperation rather than confrontation with farmers. The Guide neatly incorporates the 'business friendly enforcement procedures' set out in the Deregulation and Contracting Out Act 1994,[78] echoing the sentiments of the Scott Report some fifty years earlier.

In the long wait for legislation in this area, some individuals responded to the loss of hedgerows by creatively using existing law. In *Seymour* v *Flamborough Parish Council* (1997),[79] this law hailed from the eighteenth century enclosure movement. The subject of the dispute, which caught the collective legal imagination because of the remarkable successes of its legally untrained protagonist, was a 56 yard stretch of ancient hedge, threatened by the Council's plans to uproot it to extend a bowling green. Cracknell J. ordered the Council to preserve the 'undistinguished, badly maintained, straggly and unkempt' hawthorn hedge on the grounds that it was still bound by the Flamborough Enclosure Act 1765 which required the parish council 'to maintain forever' the hedge. This successful reliance on a local Enclosure Act demonstrates an alliance between the protection of property interests (the broad aim of the local Act) and conservatory interests, mediated through law, and in this case spanning time and purpose. The judge stressed that the judgment did not set a precedent and that each of the thousands of eighteenth and nineteenth century Enclosure Acts must be judged on their merits. But he underlined the general principle that 'the courts cannot and do not strike down a statute merely because it is old or passed by a Parliament that was very far from being elected by universal suffrage'. The implication of the judgment is that protection may be granted to hedgerows which are the subject of arcane Enclosure Acts. But those which predate such Acts still require legal protection, a point emphasised by the judge in his consideration of the then forthcoming Hedgerows Regulations.

Law and landscape

That law has played a significant part in the construction of the landscape is seen in the case of the hedgerow. Early law legitimated a great shift in the economic structure of landowning and later provided the mechanisms by which planned and regularised enclosure could take place. The particular legal form chosen for enclosure shaped the physical landscape. The hedgerows planted to give effect to the Enclosure Acts are physical representations of the social and legal history of enclosure, particularly patterns of land ownership, and rejection of common land rights. Further examples of law bearing upon landscape are the Game Laws of the eighteenth century[80] and the Corn Laws of the nineteenth century. Today, different legal landscapes can be seen: the Private Trust landscape, the Common Agricultural Policy landscape and the Set Aside landscape, all of which are structured and upheld by legal instruments, concepts and doctrines.

Although law has shaped the physical form of land, the legal protection of landscape is tenuous. The legislative measures remain oblique, setting out duties of taking the conservation of amenity and beauty of land into account in decision

78 Deregulation and Contracting Out Act 1994, s 5 and Sched 1.

79 Not yet reported, see *The Times*, 3 January 1997.

80 D. Hay *et al*, *Albion's Fatal Tree. Crime and Society in Eighteenth Century England* (London: Penguin, 1988).

© The Modern Law Review Limited 1999

making, but imposing few specific legal obligations. The recent legislative attempts at hedgerow protection pursue a one-dimensional view of landscape as no more than the sum of its surface features, so that measurement, rather than memory or understanding, is the absolute arbiter of value. Landscape is, in this case, legally understood in scientific, objective and narrowly historical terms, less so in social, popular, or cultural terms. Law therefore fails to capture the richer, extended vision of landscape – the 'inner landscape' of perception and meaning. Doubtless one reason for this is the practical inability for law to capture the importance of such incalculable concerns using legal methods which are sufficiently objective to be workable, but what is really at stake is a legal indifference to the creation of landscape by culture. Unlike in the tradition of critical geography, law has still to appreciate that landscape, like space, is in part socially constructed and to reflect this in its representations.

In the case of the hedgerow, law's contribution to the construction and protection of landscape is asymmetrical. To some extent this may be explained on the basis that law's role in shaping the physical form of land via the Enclosure Acts was primarily a consequence of its creation, then upholding, of private property rights.[81] The planting of hedgerows was a physical representation of the imposition of legal barriers to common ownership. This prevails today so that English law generally protects the material value of property. In contrast, the conservation of landscape relies in no small part on the successful subversion or at least transcending of private property, and specifically developmental interests.[82] This is most unlikely to occur for the purposes of nature conservation, although examples exist of the creative use of law by individuals such as in *Seymour* v *Flamborough Parish Council* by which a sense of history and place are captured in the legal mechanism of a local Enclosure Act so as to have relevance today.

The legal recognition of hedgerows as 'commons' may be furthered in the future by European environmental law. For example the EC Habitats Directive[83] imposes a duty on Member States to encourage the management of 'the traditional systems for marking field boundaries', or 'green corridors', which are deemed essential for the migration, dispersal and genetic exchange of wild species.[84] Such features have been described by the Court of Justice as forming part of the 'common heritage of the Community',[85] and thus capable of subjugating the subjective or national interest element of land use to 'common responsibilities'.[86] This development of a concept of a common *natural* heritage parallels that of European citizenship, raising similar difficulties of definition, competence and enforcement.[87]

---

81 See similarly with regard to the Black Acts which structured relationships in rural areas and forests, E.P. Thompson, *Whigs and Hunters: The Origin of The Black Act* (London: Allen Lane/Penguin Books, 1975).

82 See S. Whatmore, *Possessing the Earth. Property and the Environment in Late Modernity* (London: Routledge, 1995); but also E. Brubaker, *Property Rights in the Defence of Nature* (London: Earthscan, 1995).

83 Council Directive 92/43/EEC on the conservation of natural habitats and wild fauna and flora, OJ 1992 L 206/7, 21 May 1992.

84 *ibid* article 10(2).

85 Case C-44/95, *R* v *Secretary of State for the Environment, ex parte Royal Society for the Protection of Birds* ('*Lappel Bank*') [1996] 3 CMLR 411, para 24

86 See, for example, Opinion of Advocate General Fennelly in Case C-44/95, '*Lappel Bank*' [1996] 3 CMLR 411, paras 72–75.

87 See, for example, N. MacCormick, 'Democracy, Subsidiarity, and Citizenship in the European Commonwealth' (1997) 16 *Law and Philosophy* 331, 340–342

© The Modern Law Review Limited 1999

In the absence of law which is capable of appreciating landscape beyond the sum of its surface features, as the marks of property, greater reliance might perhaps be placed on the actions of individuals such as in *Seymour* v *Flamborough Parish Council*, or, at the other end of the legal spectrum, the further development of a European natural 'commonwealth'. Even more imaginative is the current grassroots plan to plant a 150 mile, cross-Britain hedgerow (already known as The Great Hedge), a type of natural Cristo artwork, across private property, common lands, and linking existing hedgerows.[88]

88 Organised by Plantlife, Natural History Museum, see J. Theolbald, 'Getting Close to the Hedge' *The Guardian*, 22 January 1993

© The Modern Law Review Limited 1999

CASES

Court-Authorised Caesarean Sections – The End of a Trend?

*Sabine Michalowski**

In recent years, several courts have been approached by hospitals asking for a declaration that it would be lawful for them to perform a Caesarean section on a patient without the patient's consent. Even though the facts of the various cases differ, almost all of these cases involved emergency hearings and required instant decisions, as in most of these cases the life of the patient and/or that of the foetus was at risk if the Caesarean section was not performed within hours of the application to the court. Before the recent decision of the Court of Appeal in *St. George's Healthcare NHS Trust* v *S*[1] in every case so far decided by a British court, the pregnant woman's refusal to consent to the Caesarean section had been disregarded and the performance of a Caesarean section without the patient's consent or even despite her explicit refusal had been declared to be lawful. In all cases but one,[2] the decision rested on the court's assessment that the patient lacked the competence to validly refuse to have a Caesarean section.[3] The urgency of the situation hardly left the courts sufficient time adequately to explore the factual as well as the legal issues arising in the particular case. Decisions in this area tend to demonstrate that the courts, when having to decide an urgent matter of life and death, prefer to decide in favour of preserving life, and the easiest way to reach such a decision is to declare the patient's incompetence. Given the far-reaching implications of this approach for the autonomy of pregnant women, the decision of the Court of Appeal in *Re MB*[4] in which the Court developed guidelines for future cases of court involvement in emergency Caesarean sections becomes all the more significant. *Re MB* provided an opportunity to clarify the legal principles governing the determination of a patient's competence, the relevance of those principles for women in labour, and the application of the best interests test where the pregnant woman is found to be incompetent. In addition, the Court of Appeal examined the legal status of the foetus and discussed whether the foetus had any rights or interests that must be taken into consideration when deciding cases of emergency Caesarean sections.

* Faculty of Law, University of Sheffield.

1 (1998) 3 All ER 673.

2 *Re S (Adult: Refusal of Treatment)* [1992] Fam 123.

3 See *Re L (An Adult: Non-consensual Treatment)* [1997] 1 FCR 609; *Norfolk and Norwich Healthcare (NHS) Trust* v *W* [1997] 1 FCR 269; *Rochdale Healthcare (NHS) Trust* v *C* [1997] 1 FCR 274; *Tameside and Glossop Acute Services Trust* v *CH* [1996] 1 FCR 753.

4 *Re MB (An Adult: Medical Treatment)* [1997] 2 FCR 541.

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

The facts

Ms MB, a 23-year-old woman, attended an antenatal clinic on 13 February 1997 when she was 40 weeks pregnant. An examination showed that the foetus was in the breech position, causing a risk of serious consequences for the unborn child, but involving little physical danger to the mother. According to the consultant obstetrician and gynaecologist, Mr N, because of the risk to the foetus it was the practice to recommend a delivery by Caesarean section. Mr N explained to the patient the risk to the foetus of a vaginal delivery, and she agreed to have a Caesarean section. Mr N did not discuss the method of anaesthesia. Ms MB was admitted to the hospital on 14 February. On admission, she signed the consent form for the Caesarean section but twice refused to undergo a venepuncture to provide blood samples because she was suffering from a needle phobia. On 15 February, she once more requested a Caesarean section but the hospital required blood samples. On 16 February Ms MB again signed a consent form for the Caesarean section and arrangements were made for the operation to take place. The anaesthetist visited the patient in order to insert the veneflon but the patient refused and the operation was cancelled, as Ms MB was neither prepared to allow blood samples to be taken nor to undergo anaesthesia by way of injection. She was then seen by a consultant anaesthetist and agreed to anaesthesia by mask without injection. After she went into labour, her general practitioner went to talk to her at about 7 pm and reported that she was happy to go for the operation provided she did not feel or see the needle, did not have an IV line and did not have a catheter after the operation. She again agreed to the Caesarean section at 8 pm when talking to the consultant psychiatrist Dr F. At 9 pm she went to the operating theatre and got on the table. When she saw the mask, however, she pushed it away and refused to consent to anaesthesia, and the operation was again cancelled.

At 9.25 pm the hospital sought a court order and Hollis J made the declarations at 9.55 pm. Ms MB instructed her solicitor over the phone to appeal against Hollis J's decision, and the Court of Appeal heard the appeal later the same night.

Establishing capacity of the pregnant woman

The question of the pregnant woman's capacity was the crux of most cases of emergency Caesarean sections that came before British courts, and it was also the most important problem facing the Court in *Re MB*. The Court took the opportunity to develop guidelines as to the basic principles of consent and of establishing capacity to be applied in such a situation, though the Court hastened to add that these guidelines could not be determinative, given that every such decision depended on the particular facts before the court.

Butler-Sloss LJ, who delivered the opinion of the Court, referred to well-established legal principles[5] when stating that every person was presumed to have capacity to consent to or to refuse medical treatment unless and until that presumption was rebutted, and that a competent woman had the right to refuse to consent to a medical intervention for religious reasons, other reasons, for rational or irrational reasons or for no reason at all, even though this might result in her own death. In such a case, a court did not have jurisdiction to override her decision and

5 See for example *Airedale NHS Trust* v *Bland* [1993] 1 All ER 821, at 866 per Lord Goff; *Re T (Adult: Refusal of Medical Treatment)* [1992] 2 FCR 861, at 877 per Lord Donaldson MR

© The Modern Law Review Limited 1999

the question of whether the proposed medical treatment was in the best interests of the patient did not arise.

The Court thus confirmed that the autonomy of a female patient is not reduced by the fact of her pregnancy. Rather, the same principles that apply to every competent adult also apply to pregnant women. While this conclusion seems obvious, the Court's statement was nonetheless necessary, given the holding of Sir Stephen Brown P in *Re S (Adult: Refusal of Medical Treatment)*.[6] In that case, a woman who was a 'born again Christian' refused on religious grounds to consent to have a Caesarean section. Sir Stephen Brown P did not declare her to be incompetent, but nevertheless granted the declaration sought by the hospital, partly on the grounds that the operation was in the vital interests of the patient and could therefore lawfully be performed despite the patient's refusal to give her consent to the operation. Such an argument, which never reflected English law, would now violate the principles laid down in *Re MB*.

In *Re MB*, the Court continued by saying that panic, indecisiveness and irrationality, even though they could be symptoms or evidence of incompetence, in themselves were not sufficient grounds for a finding of incompetence. Thus, the fact that the patient's decision may be seen as irrational did not in itself justify a finding of incompetence, unless the decision was based on a misperception of reality.

This clarification is very important, given that in *Rochdale Healthcare (NHS) Trust* v *C*,[7] Johnson J stated that:

> The patient was in the throes of labour with all that is involved in terms of pain and emotional stress. I concluded that a patient who could, in those circumstances, speak in terms which seemed to accept the inevitability of her own death, was not a patient who was able properly to weigh-up the considerations that arose so as to make any valid decision, about anything of even the most trivial kind, surely still less one which involved her own life.[8]

The judge reached this conclusion even though the consultant obstetrician was satisfied that the patient had the capacity to make a treatment decision. The judge's statement clearly shows that one of the reasons to hold the patient to be incompetent was that the patient seemed to accept the inevitability of her own death, a decision that did not seem rational to the judge. Surely, under *Re MB*, this would no longer constitute a valid argument, as even the fact that no 'sensible' person could have made such a decision is not in itself sufficient to justify a finding of incompetence.

In *Re MB*, the Court then considered which criteria had to be met to declare the incompetence of a patient. According to Butler-Sloss LJ, a person was unable to make treatment decisions if the requirements set out by Thorpe J in *Re C (An Adult: Refusal of Treatment)*[9] were not fulfilled, ie the patient was unable to comprehend and retain the information which was material to the decision, and the patient was unable to use the information and to weigh it in the balance as part of the process of arriving at the decision. Other factors that could completely erode the patient's competence were the temporary factors mentioned in *Re T*,[10] ie confusion, shock, fatigue, pain or drugs, but only if the person assessing the

6 *Re S (Adult: Refusal of Treatment)* [1992] Fam 123.
7 [1997] 1 FCR 274.
8 *ibid* at 275
9 [1994] 2 FCR 151.
10 *Re T (An Adult: Medical Treatment)* [1992] 2 FCR 861.

© The Modern Law Review Limited 1999

patient's capacity was satisfied that these factors were operating to such a degree that the ability to decide was absent.[11]

This constitutes another important clarification, as in *Norfolk and Norwich Healthcare (NHS) Trust* v *W*,[12] a pregnant woman who had a history of psychiatric treatment but at the time of the treatment decision was not suffering from a mental disorder, was found incompetent on the grounds that:

> She was called upon to make [a] decision at a time of acute emotional stress and physical pain in the ordinary course of labour made [it] even more difficult for her because of her own particular mental history.[13]

As was seen above, in *Rochdale Healthcare (NHS) Trust* v *C*[14] the decision that the pregnant woman was incompetent to make a treatment decision was also influenced by the fact that she was 'in the throes of labour with all that is involved in terms of pain and emotional stress'. In both of these cases, the court's perception of the pregnant woman's incompetence rested on the fact that the woman was in labour. Under *Re MB*, such a general presumption would not stand up to judicial scrutiny. Rather, a physician or a judge, when assessing the competence of a woman in labour, will have to be satisfied, on the facts of the individual case, that the labour pain is indeed so severe as to exclude decision-making capacity. Thus, one important consequence of *Re MB* is that in the future, women in labour are no longer presumed to be incompetent because of the accompanying pain and stress, but their competence will rather have to be assessed according to the same criteria applicable to all other patients who are in pain and who are frequently regarded as competent to refuse life-sustaining treatment.[15]

The main factor indicating the possibility of Ms MB's incompetence was her needle phobia. In that respect, the Court held that a compulsive phobia could stifle the patient's belief in and undermine the patient's ability to weigh the information presented to her in such a way as to exclude a true decision. Again, the Court stressed that careful scrutiny of the evidence was necessary, but it is surprising that the Court then distinguished between fear of an operation that may be a rational reason for refusing to consent to medical treatment, thus not excluding capacity, and fear that paralyses the will and therefore voids the patient's competence. It is submitted that the rationality of the patient's fear can be of no more significance than the rationality of other reasons of the patient. While it is perfectly true that fear can exclude capacity if it overrides the patient's ability to understand and weigh treatment information, the Court should not betray its own principles, as irrational fear does not necessarily negate the patient's ability to consent to medical treatment. The only criterion should therefore be whether or not the patient's fear resulted in an inability to understand, retain, believe and weigh the treatment information presented, and not whether or not the patient's fear was rational or irrational.

On the facts of *Re MB*, the Court came to the conclusion that the patient who wanted to have the Caesarean section, but was frightened of needles, was

---

11 For a discussion of this test to establish competence see Ian Kennedy, Commentary to *Re MB (Medical Treatment)*, in (1997) 5 *Medical Law Review* 317.

12 [1997] 1 FCR 269.

13 *ibid* 272.

14 [1997] 1 FCR 274.

15 See also Rosalind Ekman Ladd, 'Women in Labor: Some Issues about Informed Consent' in Helen Bequaert Holmes and Laura M. Purdy (eds), *Feminist Perspectives in Medical Ethics* (Bloomington and Indianapolis: Indiana University Press, 1992), 216, at 220; Barbara Hewson, 'Women's Rights and Legal Wrongs' [1996] NLJ 1385.

© The Modern Law Review Limited 1999

dominated by her needle phobia and unable to weigh any other considerations. This fear amounted to such an impairment of her mental functioning as to render her temporarily incompetent.

Even though the Court's reasoning seems to be supported by the facts of the case, it is nevertheless deplorable that the Court expressed its conclusion that the patient was incompetent without any careful application of the legal standards so meticulously outlined. The decision rather displays a discrepancy between lengthy theoretical statements about the principle of patient autonomy, and a very brief and general application of these principles to the case at hand. A more careful analysis of the patient's incompetence would have been particularly desirable given the tendency demonstrated by the lower courts in most of the cases involving Caesarean sections to lightly declare a woman in labour to be incompetent to make her own decisions. More emphasis should therefore be placed on the evaluation of the competence of the patient in any given case to avoid the risk of only verbally adhering to the principle of patient autonomy, while at the same time declaring the particular patient incompetent on feeble grounds, rather than allowing the patient to make an irreversible life-threatening decision.[16] In *Re MB*, the Court seemed to infer from the fact that Ms MB had consented to the Caesarean section that her needle phobia excluded her ability to make a decision that would allow her to achieve this goal. Yet Ms MB had always made it clear that her consent to the Caesarean section depended on certain conditions, the most important of which being that no needle pricks would be involved. She never consented to anaesthesia by injection, not even before 'the moment of panic' when 'the needle ... dominated her thinking and made her quite unable to consider anything else', as the Court seems to suggest. To pay heed to its own principles, the Court should have stated why Ms MB's needle phobia excluded her capacity, and this should have been explained without indirectly referring to her contradictory and thus irrational behaviour. According to the principles set out in *Re MB*, to rebut the presumption of Ms MB's competence, the Court had to show that her needle phobia was of such a nature as to exclude a real decision, but all the Court demonstrated was that to Ms MB, not to be subjected to a needle prick was more important than having a Caesarean section even if the latter enhanced the chances of delivering a healthy baby. Is it possible that the Court found the patient's attitude to desire a Caesarean section but not at the price of having to endure a needle prick so irrational as to justify the conclusion that the patient was temporarily incompetent?

The Court also considered how to accommodate the fact that in some cases of court-authorised Caesarean sections, the patients who had refused their consent to the operation had later expressed their satisfaction that their own decision had been disregarded and a healthy child had been delivered by Caesarean section. In this respect the Court held that however desirable it might be for the mother to be delivered of a live and healthy baby, this was not a strictly relevant consideration. The Court decided that even if the pregnant woman might later regret the outcome, the respect owed to her right to autonomy was more important than a protection of the woman from her own potential reactions to the consequences of her autonomous decision. This is an important consideration, as it could easily be felt that the patient's gratitude of having a healthy baby retrospectively justified any disregard of the competent woman's wishes. However, the Court of Appeal is right in dismissing this argument, as this approach could result in medical and

16 This tendency can be seen for example in *Re T (An Adult: Medical Treatment)* [1992] 2 FCR 861.

© The Modern Law Review Limited 1999

judicial paternalism. Physicians and courts could then decide that a patient, though competent, was likely to change her mind and later regret the consequences of her decision, and that it would therefore be warranted that the physician or the court disregard the woman's wishes as contrary to her best interests. Competence to make health care decisions presupposes that the patient is in a position to determine what is in her best interests. It also means that the patient will have to weigh the costs and benefits of the proposed treatment, and that she will have to accept and be able to live with the consequences of her decision. The fact that a patient whose wishes had been disregarded later adjusts to the situation created by the disregard of his/her original desire does not mean that the autonomous decision expressed at the time was wrong and coercion therefore justified; it rather merely shows that perceptions can change and that the human person is able to adjust to changing circumstances.[17]

It can thus be seen that the legal principles regarding the capacity to make difficult and even life-threatening health care decisions which at least in theory apply to pregnant women and women in labour, strongly protect patient autonomy. However, the significance of these principles can be questioned, as a decision declaring the incompetence of the pregnant woman can be easily made, and as courts do not even have to provide a detailed analysis justifying this result.

Best interests of the pregnant woman

Given that the Court decided that Ms MB was temporarily incompetent to make a valid treatment decision, the Court, in conformity with the holding in *Re F (Mental Patient: Sterilisation)*[18] applied the best interests test to determine whether or not the Caesarean section could lawfully be performed. The Court stated that the principles to be applied when considering the best interests of a temporarily incompetent patient were similar to those applicable when determining the welfare of a child, since the court and the doctors were concerned with a person unable to make the necessary decision for him/herself. However, to place the child patient and the temporarily incompetent patient on the same level is neither helpful nor appropriate. Given that in the case before the Court, the patient unable to make the necessary decision for herself was an otherwise competent woman suffering from a phobia, the determination of her best interests should not be compared to the test to be applied when dealing with a child patient, as this woman based her life on autonomous values and decisions which should be respected when making a treatment decision on her behalf. While English courts in several cases concerning child patients have for example held that the court is under an obligation to preserve the life of the patient even against the patient's wishes,[19] the same is not true for temporarily incompetent adult patients.[20]

The Court clarified that the interests to be considered when making a treatment decision on behalf of a temporarily incompetent patient were not limited to the patient's medical interests. This was an important clarification, as it can hardly be

17 Julian Savulescu, 'The Trouble with Do-Gooders: the Example of Suicide' [1997] 23 JME 108, 111.
18 [1990] 2 AC 1.
19 See for example *Re S (A Minor) (Consent to Medical Treatment)* [1994] 2 FLR 1065; *Re W (A Minor) (Refusal of Treatment)* [1992] 4 All ER 627 (CA); *Re R (A Minor) (Wardship: Medical Treatment)* [1991] 4 All ER 177.
20 See *Re T (An Adult: Medical Treatment)* [1992] 2 FCR 861.

© The Modern Law Review Limited 1999

said that the proposed treatment was in the medical interests of Ms MB, given that the gynaecological complications in her case posed almost no risk to her, while anaesthesia and a Caesarean section involve risks to the woman's life and health. The Court rather also took the woman's psychological well-being into account, when stating that it was clear on the evidence that the mother wanted this child to be born alive and that Ms MB was in favour of the operation, subject only to her needle phobia. Psychiatric evidence seemed to support the view that Ms MB was likely to suffer significant long-term damage if the child was born handicapped or died.

As, in *Re MB*, the mother wanted to be delivered of a healthy baby, it seems justified to take the psychological benefits into account. In *Norfolk and Norwich Healthcare (NHS) Trust* v *W*,[21] however, the court weighed this psychological benefit even though the woman had expressed that she did not care about the baby. Johnson J justified this as, in his view:

> [D]espite her present view about the foetus, [termination of her labour] would avoid her feeling any feeling of guilt in the future were she, by her refusal of consent, to cause the death of the foetus.[22]

It is submitted that psychological benefits should only be valid factors influencing the outcome of the best interests test if it can be established that in the individual case such a benefit is likely to accrue. When deciding what is or is not in the patient's best interests, the court should consider the patient's preferences and refrain from imposing its own value judgments on the patient.

In *Re MB*, the Court then weighed the possible dangers to the woman's well-being against the risks of anaesthesia and came to the conclusion that she would not suffer lasting harm from anaesthesia being administered to her, while the death of her child could cause her long-term psychological damage.

With all respect, the examination of Ms MB's best interests seems far too superficial. The risks of anaesthesia were not sufficiently examined. Moreover, looking at the effects of anaesthesia in isolation is hardly appropriate when determining the best interests of Ms MB. In the case of Ms MB, her needle phobia must also be put in the equation. To that effect, it would have been important to explore the effects of enforced treatment on a person suffering from such a phobia. Case studies on patients with needle phobia seem to suggest that they typically avoid seeking medical care, and this may well have been the reason why Ms MB did not regularly attend antenatal clinics throughout her pregnancy. Needle phobia seems to be increased by exposure to needles in traumatic settings,[23] and treatment to overcome the phobia must be based on trust.[24] Regard should also be had to the availability of alternatives to treatment involving needle pricks, or the possibility to help the patient overcome her phobia,[25] rather than subjecting her to coercive medical treatment. It is thus submitted that the balancing exercise as performed by the Court when determining Ms MB's best interests was unsatisfactory, as the Court did not sufficiently weigh all the risks

---

21 [1997] 1 FCR 269.

22 *ibid* at 272.

23 James M. Ferguson, C.B. Taylor, Bruce Wermuth, 'A Rapid Behavioural Treatment for Needle Phobics' (1978) 166 *Journal of Nervous and Mental Disease* 294.

24 Barbara Dennis, 'Care Study: Severe Needle Phobia' (1994) *Midwives Chronicle & Nursing Notes* 58; James M Ferguson *et al*, n 23 above.

25 See for example James M Ferguson *et al*, n 23 above, reporting the case of a woman with severe needle phobia that could be sufficiently reduced within one hour to allow her to consent to treatment involving needle pricks and to decrease her anxiety regarding such treatment in the future

© The Modern Law Review Limited 1999

involved in the forced treatment, though this is not to suggest that the outcome in *Re MB* was necessarily wrong. However, it must be noted that the Court's reference to *Re F*[26] suggests that the Court would rely on the *Bolam* test[27] when determining the best interests of the patient. The requirements of that test are in all likelihood fulfilled, as the medical evidence before the Court indicates that a responsible body of medical opinion would regard anaesthesia and the Caesarean section in these circumstances as being in the best interests of the patient.

In exclusively focusing on the risks of anaesthesia, the Court implied that the risks of the Caesarean section as such did not have to be considered, given that the patient had already consented to that procedure. Although it may at first sight seem surprising that the best interests of a patient in need of an operation can be determined by looking at the less intrusive procedure of anaesthesia, without at the same time weighing the risks involved in the far more intrusive treatment to be performed under anaesthesia, this approach is nevertheless justified. In *Re MB*, the patient who was in principle competent to make health care decisions had made an informed choice to be delivered by Caesarean section. She had therefore herself balanced the risks and benefits of the Caesarean section as such, and had come to the conclusion that the benefits of such a procedure prevailed. Respect for her autonomy required that this decision should be respected, so that there was then no room for the Court to replace the patient's judgment as to her best interests with its own. However, given her needle phobia, the same patient was held to be temporarily incompetent when it came to giving consent to the necessary anaesthesia. Only this decision was therefore to be made on her behalf.

It was thus not necessary to take account of the risks involved in the Caesarean section as such. It is unfortunate that the Court did not have to consider this problem, as it was at the heart of some of the lower courts' decisions, and the problem is very likely to occur in the future. In every case decided by a British court, the court was happy to follow the medical evidence that the Caesarean section was in the best interests of the pregnant woman, without so much as mentioning the risks involved in the procedure. It therefore seems important to draw attention to the fact that a Caesarean section involves invasive surgery. Apart from the risks inherent in anaesthesia, a Caesarean section poses other physical risks to the woman. Some studies for example have demonstrated that 9 to 15 per cent of Caesarean sections result in serious maternal morbidity, mainly caused by endometriosis, urinary tract infection and wound infection. Women who have a Caesarean section are also more likely to have problems during subsequent pregnancies, labours and deliveries than women with vaginal delivery. In addition, even though the number of deaths from Caesarean sections is declining, the risk of maternal mortality is between 2 and 11 times higher with a Caesarean section than it is with vaginal delivery.[28] Caesarean sections may also have psychological effects on the woman, such as depression and feelings of guilt, and can adversely affect the mother-child relationship.[29] It is therefore submitted that courts, when deciding whether it is in the best interests of a woman to have a Caesarean section, should carefully balance the risks involved against the potential benefits. In

26 *Re F (Mental Patient: Sterilisation)* [1990] 2 AC 1.

27 *Bolam* v *Friern Hospital Management Committee* [1957] 1 WLR 582.

28 Helen Churchill, *Caesarean Birth, Experience, Practice and History* (Cheshire: Midwives Press, 1997) 93–97

29 Helen Churchill, n 28 above, 89; Sarah Clement, *The Caesarean Experience* (London: Pandora Press 1991) 93–110

© The Modern Law Review Limited 1999

addition to a thorough analysis of all risks possibly ensuing from a Caesarean section, it seems essential that alternatives to the treatment proposed by the physicians and rejected by the patient be fully explored.[30] This involves the question whether the Caesarean section is in fact essential for the delivery of a healthy baby, given that medical opinion seems to be far from unanimous with regard to the need for a Caesarean section, particularly if the foetus is in the breech position.[31]

Another aspect to be considered in this context is the effect of forced treatment.[32] In *Re AC*,[33] an American case, it was pointed out that court-ordered Caesareans can erode the element of trust between pregnant women and their doctors and drive women at high risk of complications during pregnancy and childbirth out of the health care system.

Given the importance of the decision whether or not a Caesarean section is in the best interests of the pregnant woman, a careful balancing of all the interests at stake seems necessary. A comprehensive analysis of the patient's interests requires knowledge of the medical procedures involved, but also recognition of the possible psychological and social effects the decision may have on the patient.

Impact of the interests of the foetus

So far, the discussion has exclusively focused on the pregnant woman's autonomy and interests. However, some of the lower court decisions on Caesarean sections, were explicitly or at least covertly influenced by concerns about the best interests of the foetus. In *Re S (Adult: Refusal of Medical Treatment)*,[34] Sir Stephen Brown P openly considered the best interests of the foetus when granting the declaration sought by the hospital on the grounds that the operation was vital to protect the life of the unborn child and could therefore be lawfully performed despite the patient's refusal to give consent to the Caesarean section. Sir Stephen Brown P referred to the statement made by Lord Donaldson, MR in *Re T*[35] that:

> An adult patient who, like Miss T, suffers from no mental incapacity has an absolute right to choose whether to consent to medical treatment, to refuse it or to choose one rather than another of the treatments being offered. The only possible qualification is a case in which the choice may lead to the death of a viable foetus. That is not this case and, if and when it arises, the courts will be faced with a novel problem of considerable legal and ethical complexity.[36]

In the light of the decision in *Re S (Adult: Refusal of Medical Treatment)*,[37] the Court of Appeal in *Re MB* discussed the question whether the competent mother's refusal could be disregarded in the best interests of the foetus, even though (because of the finding of incompetence) the question did not arise in *Re MB*. It should be noted, however, that a related question could also arise if the woman was

30 Michael Thompson, 'After *Re S*' [1994] *Med L Rev* 127, 135.
31 See Helen Churchill, n 28 above, 72 for further references.
32 Kristina Stern, 'Court-Ordered Caesarean Sections: In Whose Interests?' (1993) 56 MLR 238, 241
33 *Re AC* [1990] 573 A 2d 1253 (DC CA).
34 [1992] Fam 123
35 [1992] 2 FCR 861.
36 *ibid* 865.
37 [1992] Fam 123.

© The Modern Law Review Limited 1999

found to be incompetent, as the outcome of the best interests test could differ, depending on whether the court would have to concentrate exclusively on the best interests of the patient, or whether the interests of the foetus would also have to be taken into account.

In *Re MB*, the Court cited English authority such as *Paton* v *British Pregnancy Advisory Service Trustees*[38] and *Re F (In Utero)*[39] to the effect that the views of Lord Donaldson MR and of Sir Stephen Brown P were contrary to English law, and summarised the current position as follows:

> Although it may seem illogical that a child capable of being born alive is protected by the criminal law from intentional destruction, and by the Abortion Act from termination otherwise than as permitted by the Act, but is not protected from the (irrational) decision of a competent mother by refusing medical intervention to avert the risk of death, this appears to be the present state of the law. ... A competent woman who has the capacity to decide ... may refuse to consent to the anaesthesia injection in the full knowledge that her decision may significantly reduce the chance of her unborn child being born alive. The foetus up to the moment of birth does not have any separate interests capable of being taken into account when a court has to consider an application for a declaration in respect of a Caesarean section operation. The court does not have the jurisdiction to declare that such a medical intervention is lawful to protect the interests of the unborn child even at the point of birth.[40]

Thus, the competent woman's refusal to consent cannot be overridden in the interests of the foetus, and when applying the best interests test, only the interests of the pregnant woman, but not the interests of the foetus can be taken into consideration. The holding in *Re MB* is not only important in that it overruled *Re S*[41] – some of the lower courts' decisions, while not questioning the principle that the foetus does not have any rights, show the courts' uneasiness at disregarding the interests of the foetus. In *Norfolk* and *Norwich Healthcare (NHS) Trust* v *W*,[42] for example, the judge's final statements were that:

> Throughout [my] judgment I have referred to 'the foetus' because I wish to emphasize that the focus of my judicial attention was upon the interests of the patient herself and not upon the interests of the foetus which she bore. However, the reality was that the foetus was a fully formed child, capable of normal life if only it could be delivered from the mother.[43]

This is an illuminating statement, as it shows that Johnson J was in fact concerned about the interests of the foetus. It therefore seems important to highlight that in cases of emergency Caesarean sections, the foetus' interests cannot be taken into consideration, let alone override those of the pregnant woman. Given the courts' tendency to find a way around legal principles perceived as undesirable, illogical etc by interpreting the facts of the individual case in such a way as to allow them to achieve a desirable and logical result, it seems important to appreciate that the legal principles outlined in *Re MB*, contrary to the Court's suggestion, are neither illogical nor inconsistent. The protection of the foetus by the provisions of the Abortion Act 1967 and of the Infant Life (Preservation) Act 1929 does not go beyond protection against intentional destruction. However, the pregnant woman's refusal to consent to a Caesarean section cannot be seen as the intentional destruction of the foetus. The

38 [1979] QB 276
39 [1988] FCR 529.
40 [1997] 2 FCR 541, 558–561
41 [1992] Fam 123
42 [1997] 1 FCR 269.
43 *ibid* 272–273

© The Modern Law Review Limited 1999

patient's refusal is nothing other than the exercise of her right to make autonomous health care decisions and to protect her bodily integrity from unwanted surgical procedures.[44] This is no more an intentional destruction of the foetus than the refusal of treatment necessary for preserving the life of the woman herself would be regarded as suicide.[45] In both cases, death would be caused by nature taking its course, and all the pregnant woman could be accused of would be that she did not take any active measures to preserve the life of the foetus. However, neither the provisions of the Abortion Act 1967, nor those of the Infant Life (Preservation) Act 1929 impose on the pregnant woman the duty to actively save the foetus' life. The legal principles referred to by the Court are therefore not inconsistent, but rather regulate different situations.

On appeal in *Re AC*,[46] the ruling of the first instance court, that a Caesarean section could lawfully be performed on a woman who was close to dying from cancer and refused to consent to the operation, was rejected. The essential principle was that persons should not be compelled to accept invasions to their bodily integrity for the sake of another person's health. Moreover, even though a pregnant woman who had freely chosen to lend her body to bringing a child into the world might be argued to have enhanced duties in relation to the foetus, the foetus could not have rights that went beyond those held by persons who were already born.

This argument equally applies to English law, as no duty to rescue third parties is recognised. In the particular case of a conflict between the interests of the pregnant woman and those of the foetus she is carrying, May LJ remarked in *Re F (In Utero)*[47] that:

> [T]o accept such jurisdiction [over the foetus] and yet to apply the principle that it is the interest of the child which is to be predominant is bound to create conflict between the existing legal interests of the mother and those of the unborn child and ... it is most undesirable that this should occur. ... [T]here would be insuperable difficulties if one sought to enforce any order in respect of an unborn child against its mother, if that mother failed to comply with the order. I cannot contemplate the court ordering that this should be done by force nor ... by committal.[48]

Therefore, no competent pregnant woman can be forced to have a Caesarean section against her will, even if her refusal will cause the death of the foetus, as the woman is under no legal obligation to rescue the foetus.[49] Equally, no incompetent pregnant woman can be subjected to a Caesarean section where this operation is not in her, but rather only in the foetus' best interests. If the foetus were awarded interests other than the interest not to be intentionally destroyed if none of the exceptions of the relevant Acts apply, and if those interests could prevail over those of the pregnant woman, it would follow that the state could control the life of pregnant women in the name of the protection of the foetus.[50] Instead of respecting the pregnant woman's autonomy, such a concept turns the woman into a means for a third party's, the foetus', ends, thereby undermining not only the pregnant

---

44 *St. George's* case, n 1 above, 688 per Judge LJ

45 See for example *In the Matter of Claire Conroy* [1985] 486 A 2d 1209, per Schrieber J.

46 [1990] 573 A 2d 1253 (DC CA).

47 [1988] Fam 122.

48 *ibid* at 138.

49 See also Lisa Miller, 'Two Patients or One? Problems of Consent in Obstetrics', (1993) 1 *Medical Law International*, 97 109; Aurora Plomer, 'Judicially Enforced Caesareans and the Sanctity of Life' (1997) *Anglo-American Law Review* 235, 248–250.

50 Dawn E. Johnsen, 'The Creation of Fetal Rights: Conflicts with Women's Constitutional Rights to Liberty, Privacy, and Equal Protection' (1986) 95 *Yale Law Journal* 599, 619.

© The Modern Law Review Limited 1999

woman's autonomy, but also her human dignity. If women could be forced to have Caesarean sections in the name of the protection of the foetus, it would also follow that the interests of the foetus would be of higher value than those of the pregnant woman, a concept that is difficult to justify. Given the dangers involved in Caesarean sections, awarding the life of the foetus priority means risking the life of the pregnant woman, as every Caesarean section involves a slight risk of death.[51] In *St. George's Healthcare NHS Trust* v *S* the Court of Appeal has now also adopted the view that even if the foetus had interests, their protection could not justify an unwanted medical intervention in a competent woman.[52] It should be added that the guidelines from the Royal College of Obstetricians and Gynaecologists entitled 'A Consideration of the Law and Ethics in Relation to Court-Authorised Obstetric Intervention' also take the stance that competent women should not be forced into having a Caesarean section, when stating that:

> It is inappropriate, and unlikely to be helpful or necessary, to invoke judicial intervention to overrule an informed and competent woman's refusal of a proposed medical treatment even though her refusal might place her life or that of her foetus at risk.

Conclusion

The legal principles applicable when cases of Caesarean sections come before the courts are mostly satisfying, as they confirm that a pregnant woman has the same rights any other adult person enjoys. As long as she is competent to make her own health care choices, her autonomy will be respected even where this choice will put the life of the woman herself or that of the foetus at risk. In theory, therefore, *Re MB* has done a lot to promote the autonomy of pregnant women. Another question altogether is that of what consequences this decision will have in practice. As the decision in *Re MB* demonstrates, patient autonomy is not sufficiently protected if the courts promote it only verbally, while at the same time being prepared to declare a patient's incompetence without a detailed analysis of the facts of the individual case.[53] The current practice of the courts gives rise to concern that such decisions, instead of being based on a careful application of the legal principles intending to promote patient autonomy, are rather oriented towards the courts' view of what is best for the patient. Given the scarcity of cases in which the competence of adult patients to validly refuse medical treatment has been questioned by courts, one might detect some judicial bias towards the incompetence of women in labour. This is particularly obvious when contrasting the decisions reached in cases of Caesarean sections with the outcome of the analysis of competence in *Re C*,[54] a case in which the patient was in Broadmoor and was suffering from schizophrenia, but was male and not pregnant. Even though medical evidence suggested that the patient's refusal of treatment would in all likelihood lead to the patient's death, Thorpe J came to the conclusion that the patient was competent to make such a far-reaching decision. A reason for the courts' readiness to disregard the pregnant woman's choice could be that they want

51 Nancy Rhoden, 'The Judge in the Delivery Room: the Emergency of Court-Ordered Caesareans', (1986) 74 Cal LR 1951, 2002–2003.

52 n1 above, 688, 692 per Judge LJ.

53 See also Ian Kennedy, Commentary to *Re MB (Medical Treatment)* in (1997) 5 *Medical Law Review* 317, 323.

54 [1994] 2 FCR 151.

© The Modern Law Review Limited 1999

to protect the foetus, or even the pregnant woman herself, from decisions perceived as unreasonable, even though there is no legal basis for doing so. This dilemma was highlighted by Thorpe LJ who stated that:

> Whatever emphasis legal principle may place upon adult autonomy with the consequent right to choose between treatments, at some level the judicial outcome will be influenced by the expert evidence as to which treatment affords the best chance of the happy announcement that both mother and baby are doing well. Unless the recognition of this consideration is legitimated there is an obvious risk of strained reasoning.[55]

It is submitted that as long as courts can easily disregard a pregnant woman's health care decision that endangers the foetus by declaring the woman incompetent, autonomy, though in theory fully protected, is subject to judicial approval of the patient's decision.[56] Emergency hearings in which the pregnant woman will frequently not be represented adequately and in which there is not much time for the court to gather comprehensive evidence, almost necessarily result in a decision that is biased towards the opinion of the treating physicians who are most likely to provide the bulk of the evidence available to the court. In *Re MB*, the Court emphasised the need to identify potential problems as early as possible so as to avoid situations in which emergency hearings become necessary. However, even though some problems can be identified in advance, this does not necessarily mean that most cases in which physicians recommend a Caesarean section are susceptible to advance decision-making. It will frequently not be possible to anticipate every problem that might possibly occur. A general discussion of the pregnant woman's attitude towards a Caesarean section, however, will often not be helpful in ruling out later problems, as her attitude will depend on the circumstances of the individual case which will frequently only materialise during labour. It is therefore all the more important that judges are alerted to the problems of emergency hearings, that the pregnant woman is represented in court, a requirement that was emphasised by the Court in *Re MB*, but often disregarded in the past, and that clear legal guidelines exist to promote the pregnant woman's autonomy in such a difficult situation. In December 1997, the Lord Chancellor's Department issued a Consultation Paper, 'Who Decides? Making Decisions on Behalf of Incompetent Adults', indicating that Government sees a need for a reform of the law to clarify the legal principles in this area.

However, legal principles alone cannot provide an adequate safeguard for the respect of patient autonomy. As long as there is a widespread belief that the foetus should receive some protection against the pregnant woman's choices, that it is best for a woman to be delivered of a healthy child at all costs, and that a woman in labour is not competent to make valid health care decisions, more situations will occur in which legal principles will be circumvented to achieve this protection.

55 Lord Justice Thorpe, 'The Caesarean Section Debate' [1997] *Fam Law* 663, 663–664.

56 For a fundamental critique of the consent-based approach, see John A. Harrington, 'Privileging the Medical Norm: Liberalism, Self-Determination and Refusal of Treatment' (1996) 16 *Legal Studies* 348.

© The Modern Law Review Limited 1999

Pre-natal Injury and Homicide Following *Attorney-General's Reference (No 3 of 1994)*

Stephen Gough*

Any society that wishes to distinguish abortion from homicide must draw a (more or less arbitrary) line between *people,* who qualify for the full panoply of legal rights, and *foetuses*, which receive reduced protection. Rarely does the distinction cause practical problems for criminal lawyers,[1] but one recent exception has generated less comment than it may deserve. The factual background is as simple as it is tragic. Following a quarrel with his girlfriend, who he knew to be nearly six months pregnant with their child, the defendant stabbed her several times in the face, back and abdomen with a long-bladed kitchen knife. She went into labour two weeks later, giving birth to a grossly premature baby girl. While the baby had, it turned out, been wounded during the original attack, the only provable threat to her health resulted from extreme prematurity. After battling for some four months she eventually contracted a lung condition and died. A murder charge failed at trial, the judge ruling that neither *actus reus* nor *mens rea* for a homicide offence was present. The Attorney General referred that point to the Court of Appeal.[2]

The Court of Appeal

The Court of Appeal, while wary of relying too heavily on authorities predating the 1957 Homicide Act, accepted Coke's view that a person was not 'in being' until the moment of birth.[3] Yet there could be no doubt that the defendant had – if he had caused the baby's death at all[4] – caused the death of a 'person in being'; it was irrelevant that the fatal causal sequence had been set in motion before birth.[5]

* Faculty of Law, University of Southampton. Thanks are due to Julie Dickson, John Gardner, Andrew Halpin and Nick Wikeley for their very helpful comments on earlier drafts.

1 Early authorities are limited to the (far from unanimous) writings of legal scholars. The balance of opinion favoured Coke's view that 'If a woman be quick with childe, and by a potion or otherwise killeth it in her wombe; or if a man beat her, whereby the childe dieth in her body, and she is delivered of a dead childe, this is a great misprision, and no murder, but if the child be born alive, and dieth of the potion, battery or other cause, this is murder: for in law it is accounted a reasonable creature, in rerum natura, when it is born alive' Co Inst, Pt III, ch 7, 50. See *West* (1848) 2 Car & K 784 and *Kwok Chak Ming* v *The Queen* [1963] HKLR 349. Cases on related points are limited to *Senior* (1832) 1 Mood. 346 and *McCluskey* v *HM Advocate* [1989] RTR 182 (both on manslaughter) and *Tait* [1989] 3 All ER 682 (on threats to kill). See generally J. Temkin, 'Pre-natal Injury, Homicide and the Draft Criminal Code' (1986) 45 CLJ 414, 414–420.

2 Criminal Justice Act 1972, s 36

3 [1996] 2 WLR 412, 418. Still older authorities might have drawn the line at the point of 'quickening' when the unborn child begins to move. See Bracton Folio 121 and Fleta Book I, c23, both cited by J. Temkin in 'Pre-natal Injury, Homicide and the Draft Criminal Code', n 1 above, 414

4 *ibid*, 426.

5 Lord Taylor CJ noted the anomalies that would be created by holding otherwise: 'a defendant who poisoned the water of a pregnant woman intending her to drink it and be killed, would not be guilty of murder if the woman gave birth to a child and then made up a bottle for the baby using the poisoned water which killed the child. On the other hand, if at the time of the poison being added, the child had already been born, and the mother for whom it was intended used the poisoned water in precisely the same way with the same consequences, it would amount to murder' *ibid* 421–422.

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

Much more difficulty was presented by the defendant's mental state: murder has always been satisfied by intent seriously to injure a person in being – 'implied malice' as it is sometimes called – but any intent towards the child at the time of the stabbing must necessarily have been an intent to harm a *foetus*. While such an intent might be thought a culpable enough mental state to ground murder,[6] the Court of Appeal considered it both undesirable and unnecessary to treat it as such: undesirable for reasons unknown, although we may suspect an unexamined attachment to some version of the 'correspondence principle';[7] unnecessary because much the same result could be achieved by finding 'implied malice' towards the mother and using 'transferred malice' to attach it against the baby.

> In the eyes of the law the foetus is taken to be a part of the mother until it has an existence independent of the mother. Thus an intention to cause serious bodily injury to the foetus is an intention to cause serious bodily injury to a part of the mother just as an intention to injure her arm or her leg would be so viewed. Thus consideration of whether a charge of murder can arise where the focus of the defendant's intention is exclusively the foetus falls to be considered under the head of transferred malice as is the case where the intention is focused exclusively or partially upon the mother herself.[8]

If it is unexpected or even inappropriate[9] for the court to resort to transferred malice in a case where the victim is precisely the person intended, it is less of a surprise that it invoked a related 'unexpected mode' rule: the defendant may not have been mistaken about the identity of his victim, but he was certainly mistaken about the way in which harm would occur.

> It is clear from *Mitchell's* case [1983] QB 741 that it is unnecessary for the precise mechanism of death to be foreseen in manslaughter and we are satisfied that the same is true for murder. [...] Obviously if the mode of death is utterly remote, there may be circumstances in which this could be regarded as severing the chain of causation but in the instant case we cannot see that it should matter whether the child dies after birth as a result of a stab wound suffered by the foetus before birth or as a result of premature birth induced by the stabbing.[10]

Perhaps we should add that 'transferred malice' and 'unexpected mode' are characteristics *of* particular offences, not general or independent rules applying *to* them.[11] It is puzzling that Seneviratne criticises the appeal court's use of

---

6 See C. Wells and D. Morgan, 'Whose Foetus Is It?' (1991) 18 JLS 431; J. Keown 'Can One Murder a Foetus?' (1995) 139 SJ 914.

7 Examples of such reasoning are not hard to find in the academic literature. Temkin notes that 'For the purposes of English criminal law, a foetus is not a person. Murder requires an intention to kill or do grievous bodily harm to a person. A defendant who intends to kill or do serious injury to a foetus does not therefore have the *mens rea* for murder'. See Temkin, n 1 above, 414. See also the 7th ed J. Smith and B. Hogan, *Criminal Law*, (London: Butterworths, 1992), 329. In truth, we may rightly be sceptical about a doctrine that replicates at the level of *mens rea* a distinction that was arbitrary at the level of *actus reus*. A more rigorous analysis of correspondence can be found in J. Horder, 'A Critique of the Correspondence Principle in Criminal Law' [1995] Crim LR 759

8 [1996] 2 WLR 412 per Lord Taylor CJ, 422

9 'If a third party assaults a child during birth with intent to kill it and thereby causes it to die after birth, is it not preferable to base his liability for murder on his actual intent to kill the child rather than a fictional intent to injure the mother? (*A fortiori* if the mother herself is the assailant)': J. Keown, 'Homicide, Fetuses and Appendages' (1996) 55 CLJ 207, 209.

10 [1996] 2 WLR 412, 425 The doctrine is not new: see G. Williams, *Criminal Law: The General Part* (London: Stevens and Sons, 2nd ed, 1961) at p 125; A.J Ashworth, *Principles of Criminal Law* (Oxford: Clarendon Press, 2nd ed, 1995) 197.

11 How could we square an independent or free-standing transferred malice doctrine with offences (like 'compassing the death of the King' under the Treason Act of 1351, or wilfully obstructing a police officer under s 89(2) of the Police Act 1996) the *mentes reaes* of which are sensitive to the victim's identity? How could we reconcile a free-standing unexpected mode doctrine with offences (like rape

© The Modern Law Review Limited 1999

transferred malice on the ground that malice aforethought is inherently insensitive to the identity of the victim,[12] for an admission that malice aforethought is insensitive to the identity of the victim surely *constitutes* an admission that the doctrine of transferred malice applies to murder.

The House of Lords

Like the Court of Appeal, the Lords accepted that there was evidence upon which a jury could find the *actus reus* of a homicide offence. Lord Mustill briefly considered whether murder might require 'killing' as opposed to the more passive 'causing death', although he was prepared to accept that the latter was probably enough.[13] Nor is it uncommon for a relatively weak causal link to suffice for grave crimes, perhaps because so much of the inculpatory burden is shouldered by the defendant's evil intent.[14] The appeal court's reasoning with respect to *mens rea*, though, received short shrift. Lord Mustill thought it anatomically unacceptable to treat intent against the foetus as intent against the mother[15] – although why anatomy should be thought a fit guide to law remains a mystery[16] – and we are told that there are anyway fundamental objections to transferring malice from a pregnant woman to an ailing child after birth. Transferred and implied malice are cast as thoroughly unjustifiable vestiges of ancient legal fictions.[17] They might *just* sustain a murder conviction where the defendant intentionally stabbed a pregnant woman, inflicting injuries on the foetus which caused death *ex utero*.[18] On no account, though, should such doctrines be extended to cover the situation where there is 'malice towards the mother, the contemporaneous starting of a train of events, and the coming to fruition of those events in the death of the baby after being born alive'.[19] In fact, the only distinguishing characteristic of this second scenario is the defendant's failure to foresee the *manner* in which the ultimate harm is inflicted, and that is a *lacuna* that neither implied nor transferred malice – howsoever extended – could fill. What is *actually* required is the Court of Appeal's separate 'unexpected mode' doctrine, the existence of which we must, if we are to make any sense of the ruling at all, take Lord Mustill to be denying.

---

or assault occasioning actual bodily harm under s 47 of the Offences Against the Person Act 1861) the *mentes reaes* of which are sensitive to the manner in which the harm occurs? See further, on this last point, A.J. Ashworth's discussion of criminal damage in 'Transferred Malice and Punishment for Unforseen Consequences', in P.R. Glazebrook (ed), *Reshaping the Criminal Law* (London: Stevens and Sons, 1978) 77.

12 M. Seneviratne, 'Pre-Natal Injury and Transferred Malice: The Invented Other' (1996) 59 MLR 884, 889.

13 [1997] 3 WLR 421, 437. Compare Lord Hope who had 'no difficulty in finding in the facts of this case all the elements which were needed to establish the *actus reus* both of murder and of manslaughter', 440.

14 cf J. Gardner, 'Rationality and the Rule of Law in Offences against the Person' (1994) CLJ 502.

15 'The mother and the foetus were two distinct organisms living symbiotically, not a single organism with two aspects. The mother's leg was part of the mother; the foetus was not' [1997] 3 WLR 421, 429. See also Lord Hope, 440. It is as well to add that, from this point onwards, the Lords assume a defendant who intends injury to the mother rather than to the foetus. The Court of Appeal had, of course, assumed a defendant who intended injury to the foetus rather than to the mother. A clearer example of the potentially treacherous character of reference cases is hard to imagine. On the proper scope of the reference procedure see Lord Mustill, 437–438.

16 One might expect legal and moral principle to *guide* science, not be guided *by* it.

17 [1997] 3 WLR 421, 432.

18 *ibid*, 429.

19 *ibid*, 429. Also 434.

© The Modern Law Review Limited 1999

If this is a correct interpretation of Lord Mustill's stance, it is indeed a puzzling one. His objections to the 'transferred' and 'implied' malice rules and his refusal to countenance the related 'unexpected mode' doctrine seem, on the face of things, to be driven by a deep suspicion of legal presumptions. Yet while such doctrines certainly *are* presumptive, it is a mistake to think that the *mens rea* of murder would be any *less* presumptive if we did away with them and insisted that the killer's intent should correspond perfectly with his actions. A fiction is always required to get from 'intentional killing' to 'malicious killing', for it is not true that every intentional killing is maliciously motivated.[20] Nor is murder out of the ordinary in this respect, for serious offences routinely presume disreputable character states like malice or recklessness on the basis of neutral indicators like causation, intent and belief.[21] It is quite wrong to condemn these tightly circumscribed modern rules by association with their lax early-modern ancestors.[22]

Perhaps Lord Mustill was objecting not to legal presumptions *per se*, but only to presumptions of murderous malice in circumstances where intent does not correspond *precisely* with conduct. Why he would want to advocate so demanding a variant of the correspondence principle is harder to fathom,[23] for the resulting rules would draw some extremely unappealing distinctions. It would follow, for example, that the assassin who shoots at a married couple killing the husband is brought within murder if he intends to kill the husband, or possibly to kill either,[24] but not if his sole intent is to kill the wife. No more palatable is the distinction that would be drawn between the killer who sets out to beat his victim to death and the killer who sets out to beat him to within an inch of his life, or between the gunman whose victim dies from the gunshot wound and the gunman whose victim would have done so but for a shock-induced heart attack. If anything it would make more sense for Lord Mustill to be attacking the *cumulative* application of the various presumptive doctrines:[25] the assumed defendant set out to injure a pregnant woman by stabbing and ended up killing a baby by inducing premature birth; maybe it *is* going too far to presume malice in someone who intended neither actual harm, actual victim, nor actual mode. Even if this is right, though – and many would

20 Self-defence and provocation, for example, defeat malice but not intention Tension between presumed and actual malice is also a factor in our discomfort at classifying mercy killers as murderers.

21 The alternative would be to treat malice or recklessness as a character state like 'dishonesty' or 'mischievous discretion' that the prosecution must prove from scratch in every case. Perhaps that approach has something going for it in the special context of murder, but there is no evidence that it represents Lord Mustill's view.

22 Which is just what Lord Mustill seems to do. With respect to implied malice, for example, 'The obvious recourse is to ascribe this doctrine to the last vestiges of the murder/felony rule, and to see in it a strong example of that rule, for unlike the more extravagant early manifestations it offers at least some resemblance in nature and degree between the intended act and its unintended consequences. It would follow, therefore, that when the murder/felony rule was expressly abolished by section 1 of the Homicide Act 1957 the only surviving justification for the "grievous harm" rule fell away, with nothing left' [1997] 3 WLR 421, 431.

23 Perhaps it stems from extreme subjectivism, for strict correspondence ensures that offences reflect what the defendant intended or understood himself to be doing rather than what he was actually doing. See A J Ashworth, 'Transferred Malice and Punishment for Unforseen Consequences' in P R. Glazebrook (ed), n 11 above

24 On Lord Mustill's account 'indiscriminate malice' – the mental state of a terrorist who hides a bomb on an aircraft, or of a gunman who fires at random into a crowd – conveniently turns out to satisfy the *mens rea* for murder. 'The intention is already aimed directly at the class of potential victims of which the actual victim forms part. The intent and the actus reus completed by the explosion are joined from the start, even though the identity of the ultimate victim is not yet fixed. So also with the shots fired indiscriminately into a crowd.' Per Lord Mustill [1997] 3 WLR 421, 434.

25 There is some evidence to support this interpretation of Lord Mustill's opinion See *ibid*, 429 and 435.

© The Modern Law Review Limited 1999

argue otherwise – it would nevertheless still be wrong to reason from the cumulative invalidity of the presumptive rules to their *individual* invalidity. Lord Mustill clearly commits himself to the latter position.

Having rejected a verdict of murder, it is surprising how easily the Lords felt able to fall back on the alternative of manslaughter. Relying on established authorities, Lord Hope asserted that manslaughter would be committed wherever death was caused by an unlawful act that was also dangerous. He did not consider 'dangerousness' to be particularly demanding, requiring only an act likely to injure another person.[26] Nor, in his view, need a defendant intend to cause harm to any particular person, it being enough that sober and reasonable people would recognise a risk to *someone*.[27] A requirement of foreseeable harm to the *actual* victim would, it is implied, impose an intolerable burden on the prosecution and make for extremely convoluted jury directions.[28] Why these practical difficulties did not equally merit consideration with respect to murder we are not told. Nor is it clear why the skeletons in murder's cupboard are so ruthlessly exposed to view when unlawful act manslaughter's equally murky past is barely mentioned.[29]

Conclusion

> One could expect a developed system to embody a law of murder clear enough to yield an unequivocal result on a given set of facts, a result which conforms with apparent justice and has a sound intellectual base. This is not so in England, where the law of homicide is permeated by anomaly, fiction, misnomer and obsolete reasoning.[30]

So Lord Mustill opens his speech. Yet the anomalies, fictions, misnomers and obsolete reasoning of the old cases frequently conceal important insights while the 'intellectual base' of the reformer can easily become a smoke screen for ill-considered changes. So, unfortunately, the Lords introduce more uncertainties into modern homicide law than they remove. Doubt is cast on the justifiability, if not the legal validity, of otherwise established doctrines like transferred and implied malice. A related 'unexpected mode' doctrine is either invalid *per se*, or at least invalid when used in combination with the transferred malice and implied malice doctrines; we are not sure which. Presumptive rules generally – the bedrock of a workable criminal system for centuries – are scorned without any attempt to understand their role, or any sense of what could be expected to replace them. And while the all-inclusive manslaughter offence is eventually wheeled in to save the day, we have no idea why it escapes many of the same objections. Let us hope that the Lords find an early opportunity to reconsider.

---

26 *ibid* 442.

27 *ibid* 443.

28 *ibid* 444–445

29 In fairness Lord Mustill recognises this difficulty and is open about it. As he notes, 'This is a strong doctrine, the more so since it might be said with some force that it recognises a concept of general malice (that those who do wrong must suffer the consequences of a resulting death, whether or not the death was intended or could have been foreseen) uncomfortably similar to the one rejected more than 150 years ago by the courts and commentators in the context of murder; and one which, it is proper to add, I have proposed in the first part of this speech, should be rejected once again in that context' *ibid* 436–437. But, he continues, 'To look for consistency between and within the very different crimes of murder and manslaughter is, I believe, hoping for too much One can look for a result which does substantial justice, and this is what I believe verdicts that B. was not guilty of murder and guilty of manslaughter would have achieved', 437. Sauce for the goose is not necessarily sauce for the gander, but there is surely an obligation to explain why.

30 *ibid* 424.

© The Modern Law Review Limited 1999

Directors' 'Tortious' Liability: Contract, Tort or Company Law?

Ross Grantham and Charles Rickett*

In recent years authorities in most Commonwealth jurisdictions have accepted that the imposition of personal liability on a director for damage caused to a third party while acting on behalf of the company is constrained by the doctrines of company law.[1] As the embodiment of the company the director incurs no liability unless and until he assumes personal responsibility. In *Williams* v *Natural Life Health Foods Ltd* the Court of Appeal[2] also accepted this basic premise, though the majority's application of this principle was widely thought to be wrong.[3] Factually, there was little evidence that the plaintiffs had relied upon the director and the Court's conclusion that the director must have assumed personal responsibility because the advice he gave drew on experience gained before the company was formed was quite inconsistent with the need, accepted in previous authority, to restrict personal liability in the case of small and one-man companies. The House of Lords has now reversed that decision.[4] However, in holding that the liability of directors under the developing assumption of responsibility tort does not present a special case and that the tort gives effect to a voluntary, contract-like undertaking rather than an imposed duty of care, the implications of their Lordships' decision extend well beyond the facts of the case.

The case arose out of a franchise agreement under which the plaintiffs set up a health food shop in Rugby. The franchisor was the first defendant company, formed by Mr Mistlin who had successfully run a similar shop on his own account in Salisbury. Central to the plaintiffs' decision to proceed with the venture were income projections provided by the defendant company. These projections showed favourable income streams with substantial profits after two years. The projections, however, proved to be inaccurate and the plaintiffs were forced to close their shop with substantial losses after only eighteen months of operation.

The plaintiffs' claim alleged that the projections were negligently prepared and sought damages from the company, though this claim was not pursued after the company was wound up. A claim was also brought against Mistlin, the founder and managing director of the company, alleging that he owed a personal duty of care in respect of the projections. In the High Court,[5] Langley J accepted that the projections had been prepared negligently and held the company to be in breach of

* Department of Commercial Law, The University of Auckland.

1 *Sealand of the Pacific* v *Robert C McHaffie Ltd* (1974) 51 DLR (3d) 702; *Trevor Ivory Ltd* v *Anderson* [1992] 2 NZLR 517; *Banfield* v *Johnson* (1994) 7 NZCLC 260, 496; *King* v *Milpurrurru* (1996) 136 ALR 327; *Microsoft Corporation* v *Auschina Polaris Pty Ltd* [1996–7] 142 ALR 111.

2 [1997] 1 BCLC 131.

3 See, for example, R. Grantham, 'Company Directors and Tortious Liability' (1997) CLJ 259 and D. Goddard, 'Corporate Personality – Limited Recourse and Its Limits' in C.E.F. Rickett and R.B. Grantham (eds.), *Corporate Personality in the 20th Century* (Oxford: Hart, 1998) ch 2.

4 *Williams* v *Natural Life Health Foods Ltd* [1998] 1 WLR 830.

5 [1996] 1 BCLC 288.

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

its duty of care to the plaintiffs. In respect of the claim against Mistlin personally, Langley J also found for the plaintiffs. Mistlin, although he had not directly prepared the projections, was to be taken to have assumed a personal duty of care to the plaintiffs because of his personal expertise. In the Court of Appeal[6] this finding was upheld by a majority. While approving the comments of the New Zealand Court of Appeal in *Trevor Ivory Ltd* v *Anderson*,[7] that directors would not normally be held to have assumed personal responsibility, Hirst and Waite LJJ found that Mistlin was to be taken to have assumed responsibility because 'the relevant knowledge and experience was entirely his qua Mr Mistlin, and not his qua director'.[8]

In the House of Lords, Lord Steyn, giving a judgment concurred in by the other Law Lords,[9] held that Mistlin was not personally liable. His Lordship began by noting that the 'technique adopted by English law to provide a remedy for the recovery of damages in respect of economic loss'[10] is the extended *Hedley Byrne* principle.[11] This principle requires that there be a special relationship between plaintiff and defendant, established by showing that the defendant assumed responsibility for the giving of the advice or the provision of services. Lord Steyn thus treated assumption of responsibility as synonymous with the *Hedley Byrne* principle. In applying this principle to companies, however, it was important to recognise and give effect to the separate legal personality of the company. Merely because a special relationship existed between the company and the plaintiff did not mean that the director was necessarily also liable. While careless action on the director's part might render the company liable,[12] if the director, or indeed any one who acts on behalf of another,[13] was to be made personally liable it was essential that the director have assumed personal responsibility for the accuracy of the advice or the provision of the services. In Lord Steyn's view, 'it is not sufficient that there should have been a special relationship with the principal [or company]... a director of a contracting company may only be held liable where it is established by evidence that he assumed personal liability...'.[14]

The crucial question was, therefore, whether Mistlin had assumed personal responsibility for the advice. Lord Steyn held that he had not. First, contrary to the majority view in the Court of Appeal, the mere fact that the relevant experience upon which the company gave advice was acquired by Mistlin in his personal capacity did not lead to the inference that Mistlin as director had personally assumed responsibility.[15] The assumption with which the law was concerned was a real one, which must be based on things said or done by or on behalf of the director. Crucially, and in line with the view taken by Sir Patrick Russell in his dissenting judgment in the Court of Appeal and by the New Zealand Court of Appeal in *Trevor Ivory*, this required more than a 'routine

6 [1997] 1 BCLC 131.
7 [1992] 2 NZLR 517.
8 [1997] 1 BCLC 131, 153, *per* Hirst LJ
9 Lords Goff, Hoffmann, Clyde and Hutton
10 [1998] 1 WLR 830, 834.
11 *Hedley Byrne & Co Ltd* v *Heller & Partners* [1964] AC 465.
12 Either by vicarious liability or through the process of attribution.
13 [1998] 1 WLR 830, 835.
14 *ibid* 835–837
15 In Lord Steyn's view, the Court of Appeal's point that, because Mistlin's experience was derived personally he must have assumed responsibility, 'will simply not bear the weight put upon it': *ibid* 838.

© The Modern Law Review Limited 1999

involvement'[16] by a director in the running of the company. Secondly, it had to be shown that the plaintiff had relied on the assumption. In Lord Steyn's view, the notion of reliance encompassed not merely *reliance in fact*, but also whether the reliance was *reasonable*.[17] In the present case, not only was there no clear evidence that the plaintiff had relied in fact on Mistlin's personal responsibility, but it would not have been reasonable to have done so. A plaintiff could not infer merely from the fact that the defendant owned and controlled the company that the latter had assumed responsibility.

The decision in *Williams* will undoubtedly be greeted with a huge sigh of relief by directors of all small and medium sized companies. Despite the lip service paid by the majority in Court of Appeal to the need to protect directors of such companies from personal liability, the effect of founding an assumption of responsibility simply on Mistlin's personal expertise was largely to deprive directors in practice of the benefit of that immunity.[18] The House of Lords decision in *Williams* is also, however, doctrinally important. Lord Steyn's judgment has important implications both for the conceptual nature of the emerging tort of assumption of responsibility, and for the basis upon which the tension between company law and the law of torts is to be reconciled. In respect of this latter issue, it is important to recognise that the source of this tension is not the principle of limited liability. Although in both the Court of Appeal and in Cooke P.'s judgment in *Trevor Ivory Ltd* v *Anderson*[19] it was said that to impose liability on directors in such cases would be to 'set at nought'[20] the principle of limited liability, in Lord Steyn's view what was really at stake was the principle of corporate personality. His Lordship's view is, with respect, clearly correct and will, it is to be hoped, lay to rest this misconception. The principle of limited liability protects the company's shareholders, and not the company or its officers. It thus has no bearing where a director is not a shareholder, and even where, as is common in small companies, directors are also shareholders it is far from clear why the individual's status as shareholder should foreclose the normal consequences of other capacities in which the individual acts.

Assumption of responsibility

The role and meaning of assumption of responsibility has been much debated.[21] Although relied upon by the House of Lords in cases such as *Henderson* v *Merrett*

16 *ibid* 836.

17 In doing so Lord Steyn approved the decision of the Supreme Court of Canada in *Edgeworth Construction Ltd* v *M. D. Lea & Associates Ltd* [1993] SCR 206.

18 R. Grantham, 'Company Directors and Tortious Liability' (1997) CLJ 259, 262.

19 [1992] 2 NZLR 517, 523. This view was repeated by Lord Cooke of Thorndon extra-judicially in *Turning Points of the Common Law* (London: Sweet & Maxwell, 1997) 18.

20 [1997] 1 BCLC 131, 154 *per* Waite LJ.

21 K. Barker, 'Unreliable Assumptions in the Modern Law of Negligence' (1993) 109 LQR 461; B. Hepple, 'The Search for Coherence' (1997) 50 CLP 67; P. Cane, *Tort Law and Economic Interests* 2nd ed (Oxford: Clarendon, 1996) 177–200. Lord Steyn's reference to the academic debate is also a further example of the increasing recognition by the judiciary of the part played by academics in the development of the law. While the work of leading academics has always, one suspects, influenced the law as applied by the judiciary, it is only recently that this influence has been openly acknowledged. See, generally, P. Birks, 'Adjudication and Interpretation in the Common Law: A Century of Change' (1994) 14 LS 156 (reprinted in B. Markesinis (ed), *The Clifford Chance Lectures: Volume 1: Bridging the Channel* (Oxford: OUP, 1996) ch 8) and 'The Academic and the Practitioner', 21st F.A. Mann Lecture, 26 November 1997.

© The Modern Law Review Limited 1999

Syndicates Ltd,[22] *White* v *Jones*[23] and *Spring* v *Guardian Assurance plc*[24] as a talisman of liability for economic loss, a number of matters have remained unclear. To the extent that 'assumption' implied an obligation arising from consent, rather than one imposed by the law, it was unclear why the law of tort should be enforcing obligations that were historically the province of the law of contract.[25] A closely related issue was how *Hedley Byrne* and assumption of responsibility were to be integrated into the general structure of the law of civil obligations. Despite the increasing tendency of English law to distinguish between liability under *Donoghue* v *Stevenson*[26] and *Hedley Byrne*,[27] it was nevertheless widely assumed that *Hedley Byrne* was a development of *Donoghue* and thus shared the latter's focus on a duty of care imposed by the general law.[28]

Although the relative brevity of Lord Steyn's judgment means it is unlikely to be the last word on this subject, two important and controversial conclusions can be drawn from his Lordship's comments. First, his Lordship expressly confirmed the consensual connotations of assumption of responsibility. It is the defendant's promise or undertaking, thus assuming responsibility, that gives rise to the obligation of the defendant. Indeed, Lord Steyn identified as the impetus for the development of the assumption of responsibility principle the 'restricted conception of contract in English Law, resulting from the combined effect of the principles of consideration and privity of contract'.[29] Assumption of responsibility was thus seen as serving 'an essential gap filling role' designed to supplement the law of contract, at least it seems until the 'House of Lords re-examine[s] the principles of consideration and privity of contract'.[30]

Secondly, Lord Steyn's treatment of the *Hedley Byrne* principle as essentially 'quasi-contractual'[31] also has important consequences for the structure of the law of civil obligations as a whole. As a category of obligation, when analysed alongside contract and unjust enrichment, torts is problematic.[32] While the law of contract and the law of unjust enrichment are constitutive of rights as well as providing for their direct enforcement, the law of torts tends to contain only second order rules.[33] The tort of conversion, for example, does not create the right to which the duty not to interfere relates. That primary right arises, necessarily prior to the commission of the tort, in the law of property. Lord Steyn's understanding of the *Hedley Byrne* principle as supplementary to the law of contract identifies it as a

---

22 [1995] 2 AC 145.

23 *ibid* 207.

24 *ibid* 296.

25 See, for example, K. Barker, note 21, p. 464. See also J. Fleming, 'Tort Law in a Contractual Matrix' (1995) 33 *Osgoode Hall LJ* 661.

26 [1932] AC 562.

27 This distinction rests largely on the basis of recovery for economic loss Thus, in those jurisdictions where recovery for economic loss is treated with less suspicion, the distinction is not as pronounced.

28 R.V. Heuston and R. Buckley, *Salmond and Heuston on the Law of Torts* 21st ed, (London: Sweet & Maxwell, 1996) 199–200, J. Fleming, *The Law of Torts* 8th ed (Sydney: LBC, 1992) 139.

29 [1998] 1 WLR 830, 837.

30 *ibid* citing Lord Goff's comments in *The Pioneer Container* [1994] 2 AC 324, 335 (PC)

31 We use 'quasi' to mean it is as though the undertaking or promise of the defendant satisfied the requirements for a contract. While Barker, n 21 above, is critical of the use of consent to explain liability in negligence, because it might lead to the type of confusion that for so long hampered the law of restitution, in our view liability under the *Hedley Byrne* principle is genuinely quasi-contractual

32 Generally, see P. Birks, 'Equity in the Modern Law: An Exercise in Taxonomy' (1996) 26 *Univ West Australia LR* 1, 15–16 and 'Definition and Division: A Mediation on *Institutes 3.13*' in P Birks (ed), *The Classification of Obligations* (Oxford: Clarendon, 1997) 24–26.

33 In the sense that they pre-suppose and are dependant upon a right arising apart from the wrong itself.

© The Modern Law Review Limited 1999

second order rule of the same order as conversion. More specifically, the primary right which the tort of assumption of responsibility seeks to enforce arises from the same event[34] as a contractual right: consent. In these terms, therefore, it appears that in English law a defendant's consensual undertaking of an obligation may give rise either to a response directly in the law of contract, or to a response indirectly in the law of tort.[35] The choice of which route is taken to the obligation is at present dictated by the principles of consideration and privity of contract. That this choice should be so constrained is not, however, necessarily undesirable. There is considerable utility in having a discrete law of contract that provides clear and stable rules which deal with the vast majority of cases. From this perspective, Lord Steyn's description of assumption of responsibility as a gap-filler seems appropriate.

Lord Steyn's characterisation of assumption of responsibility has two further implications. To the extent that *Donoghue* v *Stevenson* is premised on an obligation imposed without regard to the defendant's consent by the general law, that principle and *Hedley Byrne* are conceptually distinct. This differs from the view taken in jurisdictions such as New Zealand where *Hedley Byrne* is seen as merely a contextual application of an imposed duty of care.[36] This characterisation raises the further question of the nature and source of the right to which the imposed duty of care in *Donoghue* v *Stevenson* responds. While the imposed duty of care has been viewed as constitutive of the right to be free from injury, its focus on damage to property and the person suggests that, like assumption of responsibility, it may also be an indirect response to rights which have their origins elsewhere. In the case of property this right arises in the well articulated law of property, and in the case of injury to the person in an unarticulated but widely assumed common law right to the integrity of one's person.[37]

Directors and tortious liability

In terms of the applicable legal principles, careless or negligent conduct by company directors sits uncomfortably at the intersection of company law and the law of torts.[38] While company law places the liability exclusively on the artificial corporate entity, the law of torts imposes liability on the director as the actual tortfeasor. Historically, the law has sought to resolve this conflict by requiring, if

34 Professor Birks in particular has been a keen promoter of an 'event' based classification of the civil law. Thus, rights arise from events (consent, wrongs, unjust enrichment and others). The law then responds to the right with second order rights and remedies. See, most recently, P. Birks, 'Property and Unjust Enrichment: Categorical Truths' [1997] *NZ Law Rev* 623.

35 As a system, the common law seems to be marked by overlapping direct and indirect responses to events. Thus, a contract (arising from the event of consent) may give rise to a direct response (for example, an action in debt) and also to an indirect response through the wrong of breach of contract. The pattern of direct and indirect response is particulary evident when the concern is with property rights: see R.B. Grantham and C.E F. Rickett, 'Property and Unjust Enrichment: Categorical Truths or Unnecessary Complexity?' [1997] *NZ Law Rev* 668.

36 See, for example, the decisions of the Court of Appeal in *Brownie Wills* v *Shrimpton* [1998] 2 NZLR 320 and *South Pacific Manufacturing Co Ltd* v *New Zealand Security Consultants & Investigations Ltd* [1992] 2 NZLR 282.

37 M. Brazier (ed), *Clerk and Lindsell on Torts* 17th ed (London: Sweet & Maxwell, 1995) para 1–18. This is manifested in the torts of assault, false imprisonment and defamation

38 Generally, see J. Farrar, 'The Personal Liability of Directors for Corporate Torts' (1997) 9 *Bond LR* 102; J. Payne, 'The Attribution of Tortious Liability Between Director and Company' [1998] JBL 143; A Borrowdale, 'Liability of Directors in Tort – Developments in New Zealand' [1998] JBL 96.

© The Modern Law Review Limited 1999

the director is to be personally liable, some additional factor beyond the bare commission of the tortious act.[39] In some cases this factor was found in the director having 'procured' the tort[40] or having 'made the tort his own',[41] while more recently the courts have relied upon a notion of an assumption of personal responsibility.[42]

The effect of these additional requirements for personal liability has been to afford directors a considerable privilege. Unless the director has positively abandoned the shield of the company's separate personality, personal liability does not arise even where the director has physically committed the tortious act. Although rarely addressed directly, the most comprehensive explanation of this privilege to date was that given by Hardie Boys J in *Trevor Ivory Ltd* v *Anderson*.[43] In his Honour's view, a director is not personally liable for torts committed on behalf of the company because of the combined effect of the company's separate existence and the doctrinal process by which the company acts. While directors may act as the agents of the company, in appropriate circumstances the directors' actions, knowledge and intention will be directly attributed to the company and treated as if they were the acts, knowledge and intention of the company itself. The special treatment of directors thus arises because the tortious act is not that of the director, but of the company itself.[44]

In *Williams*, Lord Steyn also grounds the liability of directors upon an assumption of personal responsibility, and to this extent relied upon cases such as *Trevor Ivory*. However, in a number of crucial respects his Lordship's approach differs from that taken in previous authorities. First, it is not clear that the notion of assumption of responsibility in the context of tortious liability of directors relied upon in these authorities was the same concept as that embodied in the extended *Hedley Byrne* principle or, at the very least, that the concept had the same function. In cases such as *Trevor Ivory*, and in the judgments of the majority of the Court of Appeal in *Williams*, assumption of responsibility was essentially no more than a superadded element required only where the tort, of whatever type, was sought to be applied to company directors.[45] In Lord Steyn's view, however, assumption of responsibility is relevant because it is part and parcel of the tort itself.

Secondly, the explanation for the presumptive immunity of company directors from personal liability is now materially different from that found in earlier cases. In *Trevor Ivory* the immunity from liability arose from the internal workings of company law, in particular the rules of authority and attribution, and the determination that company law rules should prevail over those of torts law. The

---

39 Generally, see F.M.B. Reynolds, *Bowstead and Reynolds on Agency* 16th ed, (London: Sweet & Maxwell, 1996) para 9–114.

40 *Rainham Chemical Works Ltd (in liq)* v *Belvedere Fish Guano Co Ltd* [1921] 2 AC 465; *Performing Right Society Ltd* v *Ciryl Theatrical Syndicate Ltd* [1924] 1 KB 1 (CA); *Wah Tat Bank Ltd* v *Chan* [1975] AC 507 (PC); *C Evans & Sons Ltd* v *Spriteband Ltd* [1985] 1 WLR 317

41 *Mentmore Manufacturing Co. Ltd* v *National Merchandising Manufacturing Company Inc* (1978) 89 DLR (3d) 195; *White Horse Distillers Ltd* v *Gregson Associates Ltd* (1984) RPC 61; *King* v *Milpurrurru* (1996) 136 ALR 327.

42 *Trevor Ivory Ltd* v *Anderson* [1992] 2 NZLR 517. The Court of Appeal in *Williams* adopted this approach, and it is arguable that this was the approach taken by Kerr J in *Fairline Shipping Corporation* v *Adamson* [1975] QB 180.

43 [1992] 2 NZLR 517, 526–527.

44 Although dealing with a quite different question, Hardie Boys J's conception of the doctrinal relationship between the company and its directors seems consistent with the view taken by the House of Lords in *Re Supply of Ready Mix Concrete (No 2)* [1995] 1 AC 456 and by the Privy Council in *Meridian Global Funds Management Asia Ltd* v *Securities Commission* [1995] 2 AC 500.

45 J. Farrar, 'The Personal Liability of Directors for Corporate Torts' (1997) 9 *Bond LR* 102, 108.

© The Modern Law Review Limited 1999

decision of the Court of Appeal in *Williams* also proceeded on this basis, though the majority sought to temper the result with notions of individual responsibility drawn from torts law.[46] Lord Steyn, however, treats the matter as one entirely within the law of torts. Directors do not incur personal liability for negligent advice given in the course of the routine conduct of the company's business, not because they occupy a unique position or because the rules of company law protect them, but simply because the crucial elements of the relevant tort are not made out.

The nature of the claims against the director in *Williams* meant it was possible to dispose of the case simply by focusing on the elements of the assumption of responsibility tort itself. It is, therefore, unclear whether Lord Steyn's insistence that 'the issue is not peculiar to companies'[47] was intended to be a statement for general application. However, to the extent that his Lordship's judgment suggests that tortious liability should generally be imposed on company directors without modification, it is problematic. While on occasion there may be sound policy reasons why liability should be imposed on directors personally for torts committed in the course of operating the company, *prima facie*, company law doctrines must necessarily be accorded primacy. While such a claim may seem imperialistic, such primacy is inherent in the very nature of company law. The primary purpose of the set of rules which makes up company law[48] is to ensure that principles of law generally applicable, such as those of torts, are applied to a different and non-natural entity in a particular manner, which usually means that the scope of their application is limited.[49] Thus, although, necessarily, a director may be the actual tortfeasor or the individual responsible for a contract, the company law regime modifies the normal consequences of the director's actions, precisely to ensure that responsibility for, and the legal consequences of, the tortious conduct or contractual undertaking are not sheeted home to the individual. Where the company law regime applies, its essential function is to identify a different entity as the tortfeasor or contractor. To refuse to accept that these general principles are modified is not only to deny the primacy inherent in the rules of company law, but in a sense it is to deny the company's very existence. It is unfortunate, therefore, that this consequence was not further explored, especially in light of the attention given recently to the fundamental issues surrounding the integration of company law into the general law, by both the House of Lords[50] and the Privy Council.[51]

46 R Grantham, 'Company Directors and Tortious Liability' (1997) CLJ 259, 262.
47 [1998] 1 WLR 830, 835.
48 Indeed, in *Meridian Global Funds Management Asia Ltd* v *Securities Commission* [1992] 2 AC 500 Lord Hoffmann described the company as little more than the sum of the rules which constitute it.
49 H.L.A. Hart, 'Definition and Theory in Jurisprudence' (1954) 70 LQR 37.
50 *Re Supply of Ready Mix Concrete (No 2)* [1995] 1 AC 456.
51 *Meridian Global Funds Management Asia Ltd* v *Securities Commission* [1995] 2 AC 500.

© The Modern Law Review Limited 1999

REVIEW ARTICLES

Equity is Dead. Long Live Equity!

Alison Dunn*

Margaret Halliwell, **Equity and Good Conscience in a Contemporary Context**, London: Old Bailey Press, 1997, x + 151 pp, pb £19.95.

Should equity interfere in the market place? Mete out justice *inter se*? Retain a proprietary-based and a discretionary-based foundation? Be integrated into a more comprehensive and comprehensible juridical framework? Such questions, pertaining to the present and future role of equity as a component of the English legal system, readily spring from the most cursory examinations of equity's jurisdictions and of contemporary jurisprudence. Reasoned answers are not readily forthcoming, and a body of academic and extra-judicial writing has debated, discussed, celebrated and despaired of the use to which equity's doctrines and principles are being put.[1]

That equity has problems is axiomatic. These problems extend across theoretical and practical levels and have led to a conceptual, if not material, impasse in equity's development. Indeed, the present position of equity stands as something of a paradox. On the one hand, recent years have seen enormous leaps in the use and application of equity's doctrines and remedies, from the operation of the constructive trust in domestic and now predominantly commercial spheres, to the wider incorporation of equity in the still unrestricted arena of restitution.[2] Yet, on the other hand, doubts as to the efficacy of equity in the modern legal landscape abound, with question marks placed, for example, over the proper effect of equity's remedies on third parties, and the applicability of largely property-dominated doctrines to a widely (and wildly) changing society.[3] The consequence of these two seemingly contradictory circumstances is that whilst flourishing in one arena, equity has remained nonetheless constrained from development in a wider, more purposive sense.

In her book, which examines the role of conscience as an organising and interpretative principle underlying equity, Halliwell traces equity's malaise to two fundamental problems. First, there is the use of equity's doctrines in contexts divorced from those in which they first operated. One should hardly be surprised at the semantic and conceptual confusion evident in the present law, given that many

* Newcastle Law School.

1 A selection of this literature is cited in the notes below.

2 On the suitability of Equity for such expansion see, J. Langbein, 'The Secret Life of the Trust: The Trust as an Instrument of Commerce' (1997) 107 *Yale Law Journal* 165; P Millett, 'Equity's Place in the Law of Commerce' (1998) 114 LQR 214; A Mason, 'The Place of Equity and Equitable Remedies in the Contemporary Common Law World' (1994) 110 LQR 238.

3 As Millett has pointed out with more truth than jest, rather than change equity's rules 'some of the judges would have much preferred to be able to adapt society' instead. 'Equity – The Road Ahead' (1995) 9 *Trust Law International* 35.

© The Modern Law Review Limited 1999 (MLR 62 1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

of equity's principles are employed in late twentieth century circumstances without any real understanding of the context in which those principles were created, or the problems they were created to remedy.[4] The use of the constructive trust is the classic example of this increasingly egregious *faux pas*.[5] Secondly, over the centuries, and particularly since the time of Lord Eldon, equity's doctrines have become entrenched into 'rules', and despite equity's *pièce de résistance* – the great flexibility and discretion afforded to the judiciary – these rules have been applied as strictly as statutory provisions, thereby serving to defeat equity's spirit.

Unlike Pound, who argued at the turn of the century that the death of equity was predetermined once it organised itself into a system, the very thing it was meant to mollify,[6] Halliwell argues that equity's fatality can be averted by explicitly acknowledging and exploiting equity's subversive place in the legal system. An homogeneous equity subsumed within a greater legal whole may provide some advantages – perhaps certainty and clarity – but would sacrifice the essence of equity as a separate and ethical body of law. Capitalising upon equity's subversive qualities can be achieved by seeking a rational basis upon which to hinge its doctrines. Halliwell's belief is that unconscionability is the pivot of many equitable areas and that the doctrine of unconscionability could be more successfully utilised as one key principle upon which equity operates. Her main thesis is not to argue that equity can or should be unified under a single principle of unconscionability; rather, it is to show how the doctrine is used and how it could be better employed as one fundamental basis for equitable activity (p 18).

References to conscience and the prevention of unconscionable conduct have become increasingly prevalent in judicial pronouncements at all court levels and across many equitable jurisdictions in recent years.[7] Similarly prevalent is the connection between equity and unconscionability lauded in academic literature.[8] The connection is nothing new: equity has long been associated with the notion of conscience and the remedying of unconscionable conduct.[9] Across the centuries philosophers, writers and members of the judiciary from Aristotle, St Germain, St Thomas More and Fonblanque to Lord Mansfield, Maitland and Megarry, have touched upon the 'sweetness of mercy'[10] that is theoretically available in the courts of equity. Indeed, one need hardly look beyond equity's maxims which teach, *inter alia*, that equity looks to substance not to form, acts on conscience, does not suffer a wrong to be without a remedy and aids only those applicants who have done equity and come to the court with 'clean hands', to find an ostensible preoccupation with conscience and fairness.

---

4 See also P. Millett, 'Restitution and Constructive Trusts' (1998) 114 LQR 399, 407.

5 Consider, for example, the confusion in the demarcation between constructive and resulting trusts evident in *Westdeutsche Landesbank Girozentrale* v *Islington LBC* [1996] 2 All ER 961.

6 R. Pound, 'The Decadence of Equity' (1905) 5 *Columbia Law Review* 20, 25.

7 For example, *Taylors Fashions Ltd* v *Liverpool Victoria Trustees Co Ltd* [1982] QB 133; *Boustany* v *Pigott* (1995) 69 P&CR 298, and, among the more recent examples: *Credit Lyonnais Bank Nederland NV* v *Burch* [1997] 1 All ER 144; *Gillett* v *Holt* (1998) *The Times* June 18; and see also Millett LJ in *Co-operative Insurance Society* v *Argyll Stores (Holdings) Ltd* [1996] Ch 286, 304.

8 See, for example, T. Allen, 'Fraud, Unconscionability and Knowing Assistance' (1995) 4 *Canadian Bar Review* 29; D. Capper, 'Undue Influence and Unconscionability: A Rationalisation' (1998) 114 LQR 479; P. Milne, 'Proprietary Estoppel, Purchasers and Mortgagees: An Alternative Approach' (1997) 3 *Web JCLI*. Halliwell acknowledges that it was evidence of the revitalisation, one may even say 'renaissance', of the unconscionability doctrine which acted as a stimulus for her book.

9 Though not all agree that conscience and unconscionability lay at the original root of equity: see, for example, N. Underhill, *The Lord Chancellor* (Suffolk: T. Dalton Ltd, 1978) 91–95.

10 St. Germain, *Doctor and Student* (London: Scholar Press, 1970) 97.

© The Modern Law Review Limited 1999

The use of unconscionability as an underlying principle has more than its historic credentials to commend it. At a theoretical level, the association of conscience with moral values, social standards, fairness and justice serves to sustain the quintessential nature of equity or *epieikeia*. On a more practical level, the doctrine's inherent pliability enables a flexible approach to be taken not just to the remedy it affords, but also to its triggering requirements – two important characteristics in the face of a changing society. Whilst the use of the notion of unconscionability would be questioned by those who adhere to more formalism in law, the recent prominence of references to unconscionability in contemporary case law is a testament to its propriety in an age which appears to be moving towards broad principles and away from rigid definitions. But although unconscionability as a term is relatively fluid and apparently unsuited to an unequivocal definition, it is not an entirely nebulous notion. Halliwell states, albeit somewhat circularly, that 'the essence of the principle of unconscionability is that a party in either a social or commercial relationship with another person will not be allowed by equity to take unconscientious advantage of that other person' (p v). More tangibly, an 'unconscientious advantage', according to Sir Anthony Mason writing in an extra-judicial context, is an advantage which unreasonably and oppressively exploits the vulnerability of one of the parties in the relationship.[11] What is unreasonable or oppressive will, of course, depend upon the circumstances of a case and the context of the relationship between the parties, but would include those situations where one party's real or potential vulnerability, which has arisen from a physical attribute or entry into transactional or relational circumstance, is abused in a manner which extends beyond the remit of a hard bargain or mere folly,[12] and would be assessed in light of the abuser's state of mind.[13]

Lord Ellesmere pointed out when Lord Chancellor in the early 17th century that 'the office of the Chancellor is to correct men's consciences for frauds, breach of trusts, wrongs and oppressions, of what nature soever they be.'[14] It is clear that the doctrine of unconscionability retains this primary focus upon the actions and duties of the defendant, and this strikes at the root of the point endorsed by Halliwell that the concern of equity is not to do justice *inter se*, but to correct an injustice (p 41). In many of equity's jurisdictions, however, the reference to equity's explicit purpose as being the correction of an injustice remains unfulfilled. The gulf between correcting an injustice in theory and in practice could not be greater than in those areas analysed by Halliwell; viz equitable estoppel, equitable fraud, the acquisition of a beneficial interest in 'family' property and in, albeit to a lesser extent, accessory liability. Her two criticisms, that equity has become bound by an encroaching procedural approach to dispute resolution as guidance crystallises into rules, and that principles have become divorced from their context, can be levelled with increasing confidence and appositeness at each area.

In the context of undue influence, for example, one is readily struck by the complicated references to 'actual' and 'presumed' undue influence, the further classification of the latter into 'type 2A' and 'type 2B' undue influence, and the

11 Mason, n 2 above, 259

12 The requirement to go beyond the improvidence of a transaction is necessary since equity has never sought to 'mend a man's bargain': per Lord Nottingham, *Maynard* v *Moseley* (1676) 3 Swan 651, 655.

13 The case law surrounding the area of unconscionable bargains is particularly instructive because it has dealt specifically with an interpretation of unconscionable behaviour: consider *Evans* v *Llewellin* (1787) 1 Cox 333, 340; *Fry* v *Lane* (1888) 40 Ch D 312, 322; *Hart* v *O'Connor* [1985] 1 AC 1000.

14 *Earl of Oxford's Case* (1615) 1 Ch R 1, 6.

© The Modern Law Review Limited 1999

requirement of 'manifest disadvantage' which appears to apply to some categories, but not to others. In addition, the whole panoply of case law from the House of Lord's decision in *Barclay's Bank plc* v *O'Brien*[15] downwards, providing definitions, interpretation and guidance (though not necessarily in that order) is little short of bewildering.[16] No less is this so in the context of defining property rights and interests in the family and quasi-family home, where the flow of a bursting stream of some thirty years of case law from *Pettitt* v *Pettitt*[17] and *Gissing* v *Gissing*[18] through to *Lloyd's Bank* v *Rosset*[19] and *Drake* v *Whipp*,[20] with its definitions, interpretations, redefinitions and reinterpretations has failed to carry with it any semblance of order or clarity as to the requirements necessary to found an interest under a constructive or resulting trust, or through estoppel. Similarly, prior to the Privy Council decision in *Royal Brunei Airlines Sdn Bhd* v *Tan*,[21] the area of accessory liability with its grades of knowledge upon the Baden scale[22] and an untenable requirement of trustee dishonesty[23] would likewise have been found culpable. Even now, its impunity is not assured given the unhappy distinction drawn between commercial and non-commercial situations in the context of the test for actual or constructive knowledge for knowing receipt.[24] The point is surely not insignificant that the principal areas of modern equitable activity are in such substantial disarray to the extent that they are clearly not achieving their purpose on a wider level of correcting injustice.

To what extent the doctrine of unconscionability is a feasible peg upon which to hang these equitable actions is unclear, though Halliwell is persuasive in arguing that unconscionability as a principle could be better utilised. For example, using the judgment of Lord Templeman in *Boustany* v *Pigott*,[25] a Privy Council case in which unconscionability as a cause of action in unconscionable bargain cases was considered, she argues that 'undue influence, and even duress, are merely sub-species of the relief afforded by equity against unconscionable dealing' (pp 53–54). Unconscionable dealing, represented by the doctrine of unconscionability, is thus one general ground for relief which could unite this area. This argument has force. As Lord Macnaghten noted, fraud is 'infinite in variety'.[26] In the 18th century Lord Hardwicke identified five heads of fraud relieved by equity, including actual fraud,

15 [1993] 4 All ER 417

16 Consider, *inter alia*, *National Westminster Bank plc* v *Morgan* [1985] AC 686; *BCCI* v *Aboody* [1990] 1 QB 923; *Credit Lyonnais Bank Nederland NV* v *Burch* n 7 above; and, most recently, *Royal Bank of Scotland* v *Etridge* (1998) *The Times* August 27 in which the Court of Appeal sought yet again to interpret *Barclay's Bank plc* v *O'Brien* n 15 above.

17 [1970] AC 777.

18 [1971] AC 886.

19 [1990] 1 AC 107.

20 [1996] 1 FLR 826.

21 [1995] 3 All ER 97.

22 Encompassing five grades of actual and constructive knowledge: *Baden, Delvaux & Lecruit* v *Société Générale pour Favoriser le Développement du Commerce et de l'Industrie en France SA* [1992] 4 All ER 161.

23 This requirement stemmed from a formulaic application of Lord Selborne's dicta in the founding case, *Barnes* v *Addy* (1874) LR 9 Ch App 244, 251–252, that liability would arise where a third party assisted 'with knowledge in a dishonest and fraudulent design on the part of the trustees'.

24 Consider *Re Montagu's ST* [1987] Ch 264; *Polly Peck International plc* v *Nadir (No 2)* [1992] 4 All ER 769; *Eagle Trust plc* v *SBC Securities Ltd* [1993] 1 WLR 484.

25 n 7 above. Support can also be found in the more recent decision of the Court of Appeal in *Credit Lyonnais Bank Nederland NV* v *Burch* n 7 above. For comment, see A. Mason 'The Impact of Equitable Doctrine on the Law of Contract' (1998) 27 *Anglo-American Law Review* 1, 6; H. Tjio, 'O'Brien and Unconscionability' (1997) 113 LQR 10.

26 *Reddaway* v *Banham* [1896] AC 199.

© The Modern Law Review Limited 1999

unconscionable bargains and undue influence.[27] The categorisation of equitable fraud into these different heads is often little more than an artificial exercise, but which serves to disconnect relief from the principle which lies at its foundation.[28] These areas are implicitly connected via their broad association in equity with substantive fraud and disassociation from the formalism of common law fraud. A thread of unconscionability runs through all areas, albeit in differing degrees and often under the label 'public policy'.[29] Similarly, each area is consistent in its emphasis upon free will and the need for the spontaneous entry into a transaction. A common underlying principle could constructively draw the diverse areas together.

One problem with formalising unconscionability as a general doctrine, either in the specific area of equitable fraud or more broadly across a number of equitable jurisdictions, concerns its inevitable interpretation. If unconscionability concerns the taking of an unconscientious advantage, and an unconscientious advantage is something which is unreasonable and oppressive and which exploits vulnerability, how precisely are these defining characteristics to be construed in the light of factual circumstances? Sir Anthony Mason's argument has always been that certainty will evolve through the case law,[30] and the explicit recognition of unconscionability as an organising principle would allow the courts the freedom to begin this evolution. Yet, the judicial tendency to seek, expound and clarify guidelines in an attempt to understand the nature of a concept has already been seen to lead to more rigid and exclusive classifications than were often originally intended.[31] Take undue influence as one example: the problem lies not conceptually with its underlying nature, but materially in the way its requirements have been interpreted, or misinterpreted, and gradually crystallised into rules – such as the requirement of independent advice which has taken on a life of its own in recent years.[32] If the courts begin to grapple with an explicit application of a doctrine of unconscionability, with its references to unreasonableness, oppression and vulnerability, there is nothing to prevent the same material problem of interpretation occurring, and everything to suggest that it will. Further, as the guidelines emerge, the use of one principle of unconscionability across a number of equitable arenas could lead to the more fundamental problem of inconsistency. Will an act or circumstance which is accepted as being unconscionable in the arena of adjusting property rights between home-sharers be similarly unconscionable in the context of undue influence or liability for knowing receipt? If it would, it begs the question of whether it is necessary to have specific equitable jurisdictions, such as estoppel, accessory liability or equitable fraud, rather than one overarching jurisdiction for unconscionable conduct. But if the same act or circumstance does

27 *Earl of Chesterfield* v *Janssen* (1750) 2 Ves Sen 125, 155–157.

28 Halliwell notes in particular the similarities between abuse of confidence and undue influence and argues that they are not separate heads and do not require different rules, 53–54. Lord Browne-Wilkinson had questioned their separate basis in *CIBC Mortgages* v *Pitt* [1993] 4 All ER 433.

29 See for example, *Earl of Chesterfield* v *Janssen* n 27 above; *Huguenin* v *Baseley* (1807) 14 Ves 273; *Allcard* v *Skinner* (1887) 36 Ch D 145; and *Nocton* v *Ashburton* [1914] AC 932.

30 n 2 above, 255.

31 See text to notes 16–23 above. It was for this very reason that Lord Hardwicke argued for a concept of equitable fraud unburdened by definition. An unfettered discretion permits the maximum opportunity for inclusion of fraudulent circumstances into the ambit of the concept: see letter to Lord Kames, 30 June 1759, reproduced in J. Parkes, *A History of the Court of Chancery* (London: Longman Rees Orme Brown & Green, 1828) 508; and Lord Scarman in *National Westminster Bank plc* v *Morgan* n 16 above, 709.

32 See cases cited in n 15 and 16 above.

 © The Modern Law Review Limited 1999

not give rise to unconscionability across the board, and differing degrees of activity are required, it will become difficult to sustain a consistent and equitable approach to unconscionability on a wider level.

Halliwell's view is more positive: she has faith not only in the feasibility of a doctrine of unconscionability being embraced, but also in the doctrine being sustained if it is interpreted in a liberal rather than a constraining manner (p 49). Capper likewise has argued, in the context of equitable fraud, that a merged doctrine would contribute to legal certainty, and in fact would enable courts 'to develop clear and rational criteria for the resolution of these disputes'.[33] This can only be so where the courts lean towards developing and maintaining broad brush interpretative guidelines, and lean away from rigid rule-making. This will be achieved where a thorough examination of the values which underlie an area is undertaken, and overt expression given to them in a manner which allows decision-making to be consonant with principle, rather than preoccupied with pedantic rules. Unfortunately for Halliwell, it does not follow that, upon such examination, expression will be given to the doctrine of unconscionability. In the context of third party liability for assisting in a breach of trust, for example, a re-examination of the first principles applicable to the area led to a rejection of unconscionability as the basis of liability, the basis placed instead upon the accessory's dishonesty.[34] Whilst this explicit process of analysis receives Halliwell's support (p 114), it curtails the appropriateness of employing unconscionability on a broad scale. It also highlights the dilemma in Halliwell's argument: the problems she identifies as besetting modern equity dictate that unconscionability cannot work without understanding the basis of an area, but such an understanding will not necessarily lead to an acceptance of unconscionability as equity's cornerstone.

The most vociferous criticisms of using a doctrine of unconscionability as an organising principle, though not of greatest weight, are the traditional concerns with the undue exercise of discretion and the possible emergence of so-called 'palm tree justice'. These, it can be argued, would create uncertainty and injustice in the very area where certainty is needed most.[35] In the light of Halliwell's evidence of the consequences of the application of rules divorced from substance in a manner which can best be described as 'box-ticking decision-making', the availability of a discretionary approach is refreshing. As she argues in the context of adjusting property rights between cohabitees (chapter 4), the use of rules does not always provide clarity or certainty. The main problem in that particular area rests with the lapidary quality of property rules. 'Within the framework of such static concepts' Halliwell reasons, 'it has indeed been difficult for the courts to achieve results which are morally justifiable in the changed social circumstances' (p 74), and in this context at least, 'it is time that the courts went beyond manipulation of the property concepts and sought a more fundamental juristic basis for the adjustment of property rights' (p 86). In other Commonwealth countries this

33 n 8 above, 503.

34 *Royal Brunei Airlines Sdn Bhd* v *Tan* n 21 above. Lord Nicholls rejected unconscionability because he remained confused as to the nature of the concept. It is arguable that the liability settled on by the Privy Council in *Tan* is the doctrine of unconscionability in substance, though not expressed in form. The area focuses upon the state of mind of the accessory, and prevents an unreasonable advantage being taken which would exploit the vulnerability of a beneficiary under a trust either to a trustee's actions in breaching the trust, or to the likelihood of insolvency on the part of a trustee faced with remedying the breach.

35 See generally Sir Nicolas Browne-Wilkinson, *Constructive Trusts and Undue Influence* (The Holdsworth Club of the University of Birmingham, 1991); and Halliwell 17, 73–74.

© The Modern Law Review Limited 1999

has been achieved by the use of such discretionary concepts as unjust enrichment (Canada) or unconscionability (Australia), and in so doing the fear of palm tree justice has proved unfounded. Arbitrary decision-making may be damaging where it randomly dispenses individual notions of justice and varies with the length of the Chancellor's foot, but the basis of an unconscionability doctrine, at least as Halliwell argues it, is that by being a fundamental juristic basis it is principled, grounded and reasoned, not arbitrary.[36] Her belief is that the undue use of discretion and palm tree justice can be avoided where 'the optimum mode of language in equity ... [provides] a framework of reference which limits judicial discretion and yet allows the principle of conscience, understood in terms of society's current conception of justice and fairness, to operate as a vital principle in legal analysis' (p 9). Limiting the discretion of the judiciary, however, surely defeats the purpose of employing the doctrine of unconscionability and will bring greater reliance upon artificial rules. It will restrict both the ability of the judiciary to give expression to the values underlying the doctrine and its potential to evolve with changing social or commercial circumstances. If one accepts that the doctrine of unconscionability has value its operation should not require the limitation of judicial discretion.

More seriously, concern can be expressed as to the effect such flexible decision-making will have upon third parties. Halliwell views this objection as being 'devoid of merit since the basis of the equitable relief is the bad conscience of the party against whom relief is sought', and that 'it would not be consonant with the equitable jurisdiction invoked to allow an innocent party to be detrimentally affected by the decree' (p 37).[37] The problem, though, is that where third parties are involved, in practice it has not always proved so easy to delineate between the presence or absence of innocence. Bona fide purchasers are clearly immune, and it might even be argued that the risk to commercial third parties, such as banks and building societies, is inherent in the nature of the business that they have freely chosen to undertake. But the bona fide third party who fails to give value, the innocent volunteer, remains in a precarious position. One solution in these situations could be to take a *Re Diplock*[38] approach, thereby allowing the innocent volunteer to claim rateably. Even then, such an option could create more problems than it remedies, which proves that there is no easy solution to the issue of the effect of equitable doctrines upon third parties.

The criticisms of unruly discretion, palm tree justice and the interests of third parties relate specifically to the outward consequences of equitable activities, not to their fundamental basis. In questioning the suitability of unconscionability as an explicit organising principle upon which equity should operate, the place of unconscionability as one element at the heart of equity is not denied. Whether other elements define the basis of equity, however, needs to be considered, since they may affect the issue of whether the doctrine of unconscionability is the most suitable to be developed as the central principle of equitable activity.

By employing the doctrine of unconscionability, Halliwell seeks to argue that, as a body of law, equity contributes 'a key "morality" ' to the legal system (pp 1–2), and that 'perceptions of equity importing elements of good conscience or justice

36 Discretion has, as Lord Bingham commented, been 'put on a short leash', reigned in and exercised upon substantive principle not upon abstract notions of justice, fairness or conscience: 'The Discretion of the Judge' (1990) *Denning LJ* 27, 28.

37 Written in the context of estoppel, and approving the comments of Lunney, 'Towards a Unified Estoppel' (1992) Conv 239, 250. See also A. Mason n 25 above, 11–12

38 [1948] Ch 465.

 © The Modern Law Review Limited 1999

into the law rest absolutely on a conception of law as having a moral foundationalism' (p 141). This morality can be seen to translate itself into equity's concern to make parties stand by their obligations,[39] and in the fact that equity does not allow the legal system to legitimise conduct by permitting behaviour which is reprehensible. At its inception, this concern with morality as a drive for the development of equitable doctrines was predominantly the result of Christian influence.[40] For example, Barbour's argument that the ecclesiastical training of the Lord Chancellors accounted for the emergence and operation of specific performance with reference to the influence of the generally Christian concept of breach of faith (or *fidei laesio*) is well known.[41] It involved not just the idea that where a pledge of faith had been made it could not, in conscience, be broken, but also was influenced by the Catholic insistence upon the need for a wrong to be precisely put right for God's forgiveness to follow, and by the force of the moral obligation arising from the situation in which a promisor leads a promisee to alter position on the strength of a promise. The relevance of specifically Christian principles in today's pluralistic society has diminished, though they remain prevalent in the broad sense of social justice, and their demise has not affected the general moral standards of equity.

In focusing on conscience and in seeking to correct injustice, equity may clearly be said to be injecting a form of morality into the present operation of law. But to focus on conscience as a defining basis of equity and upon morality as the chief ingredient that equity contributes to the legal system is to neglect the wider and more prosaic sphere within which equity historically and currently operates. Whilst equity does display a concern for fairness, the way in which that concern is demonstrated has been influenced by a number of wider pressures not evident on the face of Aristotle's rhetoric, or Halliwell's examination of the doctrine of unconscionability out of equity's wider context. A survey of equity in earlier centuries, for example, would not have yielded so great an emphasis upon the remedying of unconscionability for morality's sake. Rather, what would have been revealed is how equity has been able to provide a mechanism to afford protection in areas which society has deemed important, and in so doing to support values which were defined principally by a landowning majority.[42] The strict settlement, the restraint upon anticipation and the protection of married women's property, catching bargains with heirs and expectants, the provisions made for 'ideots', 'lunatiks' and wards of court – all areas in which equity played a defining role – were not so much based upon conscience, compassion or justice as they were upon pragmatism. So too in the context of equity's development of the common law mortgage, where the great reforms which made it a more even-handed financial instrument were not always employed to protect the interests of the mortgagor and

39 Consider estoppel in this regard.

40 Christian influence upon the spirit of the Usury laws and the development of the mortgage is particularly striking: see *K & C Kreglinger* v *New Patagonia Meat Storage Co Ltd* [1914] AC 25, 35, 37. Note also S. Herman's argument that the relationship between a priest and God was the first example of a fiduciary relationship: 'Utilitas Ecclesiae: The Canonical Conception of the Trust' (1996) 70 *Tulane Law Review* 2239, 2247.

41 W Barbour, 'The History of Contract in Early English Equity', in P Vinogradoff (ed), *Oxford Studies in Social and Legal History* (Oxford: Clarendon Press, 1914) 150–168.

42 As counsel demonstrated by the chilling plea in *Earl of Chesterfield* v *Janssen* n 27 above, 129–130, a case concerning the dissipation of property through an unconscionable bargain: 'Our sons may at this moment be doing the same; and all we have laboured for may be gone just after our death.' See also M. Chesterman, 'Equity in the Law' in P. Troy (ed), *A Just Society? Essays on Equity in Australia* (Sydney: George Allen & Unwin, 1981) 75–79.

© The Modern Law Review Limited 1999

correct the conscience of the mortgagee as to enlarge the power base of the Court of Chancery at a time when battles with the common law courts were prevalent. It has even been argued that the development of many new and novel writs in Chancery in the 14th century, the period immediately after the Black Death, were not the result of attempts to soften the rigour of the common law, but rather to reorganise society and prevent the breaking down of order.[43]

It is, of course, axiomatic that many influences have shaped the development and impetus of equity. To a degree, conscience has been one motivation, but other influences have been and remain equally important. In this context, a defining element has been equity's cognisance of what may be termed 'value to society' – that is, the consequences and advantages which the pursuit of a particular doctrine or action can have over and above the immediate effect upon the individual parties. This may be deemed an attempt to inject morality, but is more likely to include less virtuous concerns. Equity's historical jurisdiction over wards of court stands as a classic example of how a protective jurisdiction can have not simply a value to interested sections of society which goes beyond a public interest in preventing the unconscionable treatment of infants, but also how that value can metamorphose according to the prevailing philosophy of society at any given period.[44] Similarly, in the context of the enforcement of charitable trusts, there is a material value to society, or a section thereof, over and above that afforded to the immediate beneficiaries and which goes beyond any attempt to correct injustice. In relation to trusts for the relief of poverty, for example, this value is classically perceived to be a general public interest in the alleviation of vulnerability, but it lies more in the economic and political reality of the removal of a financial burden from community resources. The circumstances surrounding the enactment of the 1601 Act of Elizabeth,[45] the judicial preoccupation with trust efficiency and the implicit recognition of taxation concerns[46] likewise lean to a more pragmatic basis upon which the charitable trust first developed and now operates.

This notion of value to society resonates also with the prevalence of equity in the commercial sphere. In the latter part of the twentieth century we place increasing reliance upon standards and acceptable levels of behaviour.[47] In society as a whole, chartermarks and performance indicators have become *de rigueur*. In the arena of

43 R. Palmer, *English Law in the Age of the Black Death 1348–1381* (Chapel Hill: University of North Carolina Press, 1993) 139–144.

44 Wardship was initially bound up with feudalism, which led to a focus upon the child as an economic asset of the feudal lord. Upon the abolition of feudalism and through to the late 19th century landowning principles held sway, and the operation of wardship leaned towards protection of a child's property from bounty hunters rather than protection of the child per se. Only from the late 19th century up to the removal of wardship to the Family division of the High Court in 1970 did Equity's focus rest specifically upon welfare oriented concerns. For comment, see A. R. Everton , 'The Changing Face of Wardship' (1968) 2 *Journal of the Association of Law Teachers* 18; J. Seymour, 'Parens Patriae and Wardship Powers: Their Nature and Origins' (1994) 14 OJLS 159.

45 Consider, for example, the economic and fiscal value for Elizabethan society of the 1601 Charitable Uses Act, which formed part of a systematic programme of Elizabethan legislation. As Francis More stated at the time in his *Reading on the Statute of Charitable Uses*: 'Public benefit was the key to the statute, and the relief of poverty its manifestation', cited after G. Jones, *History of the Law of Charity 1532–1827* (Cambridge: Cambridge University Press, 1969) 27.

46 Consider the operation of the 'personal nexus' test in the context of trusts for the advancement of education. *Oppenheim* v *Tobacco Securities Trust* [1951] AC 297; *Dingle* v *Turner* [1972] AC 601. There is a number of general equity cases the decisions of which appear reconcilable only on the basis of a covert attempt to prevent equitable rules being manipulated for taxation purposes. This is particularly marked in a comparison of *Re Oughtred* [1960] AC 206 and *Neville* v *Wilson* [1996] 3 WLR 460.

47 Consider, for example, the various Nolan Committee reports on Standards in Public Life.

© The Modern Law Review Limited 1999

equity, this trend has found particular expression in the establishment of accountability and regulation requirements in the context of trustee powers and duties, including investment policies,[48] and in the greater accountability of charitable bodies under the Charity Acts.[49] Given this prevailing context, two points may be made about equity's move into the market place. The first is that equity demonstrates a concern to maintain standards in commercial dealings which are designed to promote confidence and reliability and which are achieved through the strict prevention of unconscionable conduct.[50] Equity's fundamental advantage, argues Halliwell 'is that the concept of unconscionable conduct prescribes a minimum standard of conduct in relationships' (p 141). Having a minimum standard ensures that a focus is firmly placed on upholding a level of behaviour, which provides a semblance of security in a relatively insecure arena. Secondly, this concern to maintain standards, though ostensibly demonstrating a moral tendency in its link with unconscionability, is finely balanced with efficiency-based concerns.[51] In the commercial setting, equity's armoury is particularly attractive. The increasing complexity of the commercial world,[52] with its diverse forms of property transactions, dictates a more directive and far-reaching approach to the recovery of assets than the common law can provide. Further, as Langbein has indicated, the trust itself has commercial value in that it not only offers protection for beneficiaries in terms of trustee insolvency, but it also imposes a rigorous régime of fiduciary and general trustee duties, in addition to being structurally amenable to tax efficiency and flexible governance.[53] Protection against unconscionable conduct is present in the employment of these equitable principles and tools, but their use is balanced within the perimeters of the commercial sphere, with dicta strongly suggesting that equitable principles should not be employed where so to do would be to defeat the commercial expectations of parties, or more broadly the demands of commercial practice.[54]

These wider influences do not denigrate the prevalence of conscience as a pinnacle of equitable activity, nor do they seek to deny any future role for the doctrine of unconscionability as an organising concept. They do, however, serve to put conscience in context. A rejuvenation of the doctrine of unconscionability will be found wanting where it is applied not just without reference to the context in which it originated, but also where it is applied without recognition that whilst equity may be merciful it still knows how to manipulate. As was once observed, with more truth than perhaps was intended, 'Equity, like nature, does nothing in vain.'[55]

One of the problems which has led equity to its present position is that whilst traditionally it is said to follow the law, its development has been cut adrift from the overall tide of progression in jurisprudence. For equity to develop in a meaningful way, the recognition of an organising principle of unconscionability, if accepted, cannot take place as before in a vacuum. The survival of equity depends

48 *Law Commission Consultation Paper No 146, Trustees' Powers and Duties* (1997).
49 See Charity Acts 1992, 1993.
50 As demonstrated through equity's qualities of loyalty, fidelity, integrity and confidentiality: P. Millett n 2 above, 216.
51 In this respect, see A. Duggan, 'Is Equity Efficient?' (1997) 113 LQR 601; and I Duncanson, 'Equity and Obligations' (1976) 39 MLR 268, 270.
52 P. Millett n 2 above, 216.
53 J. Langbein n 2 above, 179. See also A Mason n 2 above.
54 See for example, Millett LJ in *Co-operative Insurance Society* v *Argyll Stores (Holdings) Ltd* n 7 above, 305; and Lord Browne-Wilkinson in *Barclay's Bank plc* v *O'Brien* n 15 above, 422.
55 *Brunner* v *Greenslade* [1971] 1 Ch 993, 1006, per Megarry J.

© The Modern Law Review Limited 1999

not so much upon re-conceptualising past mistakes, as it does upon a re-evaluation of the legal system as a whole. On a broad level, the overall shift in recent years towards substantive concerns may force a reappraisal of the traditional boundaries between law and equity. Lord Irvine, for example, has argued that incorporation of the European Convention on Human Rights may champion this movement because it will demand more explicit attention in judicial decision-making to the underlying bases of law and questions of 'morality and justifiability',[56] once exclusively within equity's remit.

Such a process is already at hand in the development of restitution. Restitution challenges the boundary between the common law and equity by merging together their comparable and compatible principles to form a more inclusive working relationship.[57] At present this coupling can be criticised for its 'mish-mash' of differing legal doctrine,[58] but at this still early stage in its development it promises to lead to an explicit appraisal of traditional legal categories.[59] Where reconsideration does take place, as it must, one needs to remain wary of the subordination of equitable principles. Equity is not a mannequin to be dressed in the clothes of the common law. This is particularly important since, as Halliwell demonstrates throughout her analysis, equity does have a continuing role to play in contemporary law. Equity's doctrines have been recognised for the value they can afford in a variety of circumstances to suit the needs of a variety of interests. Halliwell states that 'Fundamental misconceptions of equity abound ... because of a persistent refusal to acknowledge that equity is, by its very nature, subversive of the law' (p 6). This quality of subversion is an important asset, and should not be stifled into homogeneity. In the context of the doctrine of unconscionability, the last twenty or so years have demonstrated how a doctrine can be resurrected and applied in new and forward-thinking contexts while still retaining something of its originality. But as the greatest example of resurrection teaches, death is a necessary component. Equity's fatality is practically complete. As Halliwell clearly shows in her analysis, equity's principles are stagnating, cut off from the very lifeline that fed their birth. This is in direct contradistinction to its Antipodean cousins, whose more liberal approach has embraced unconscionability as a specific underlying basis.[60] If equity is alive and well and living in Australia, it is not inconceivable that a new equity born of well reasoned analysis and grounded in first principles may be resurrected in these colder climes. What remains to be debated is the exact form such resurrection will or should take. Halliwell's exposition of the role of conscience and the advantages of the doctrine of unconscionability in equity is a good place to begin the discussion.

56 'The Development of Human Rights in Britain under an Incorporated Convention on Human Rights' [1998] *Public Law* 221, 235; cf Sir John Laws, 'The Limitations of Human Rights' [1998] *Public Law* 254.

57 Whilst equity's notion of unconscionability and restitution's notion of unjust enrichment have some similarities, Halliwell rightly argues (at 142–146) that the distinction between the two principles should not be eroded in the present development of the law, at least not until the jurisprudence of unjust enrichment is more sufficiently developed and understood.

58 See S. Hedley, 'Ten Questions For "Unjust Enrichment" Theorists' (1997) 3 *Web JCLI*.

59 Professor Beatson in his inaugural lecture *Has the Common Law a Future?* (Cambridge: Cambridge University Press, 1997) 14, emphasised the vitality of the common law, and the contribution it could make along with equity in the context of restitution to the process of seeing how law fits into a wider system.

60 Classically represented in *Commercial Bank of Australia* v *Amadio* (1983) 151 CLR 447.

© The Modern Law Review Limited 1999

REVIEWS

Margaret F.A. Otlowski, **Voluntary Euthanasia and the Common Law,** Oxford: Clarendon Press, 1997, xxxiii + 564 pp, hb £70.00.

Support for euthanasia is in principle increasing, and the movement for legalisation of the practice is gaining momentum. Margaret Otlowski's *Voluntary Euthanasia and the Common Law* is a timely and significant contribution to this troublesome subject.

Otlowski selects five countries for comparative legal analysis: Australia, New Zealand, Canada, the United States and the United Kingdom. Her main position, made explicit from the outset, is that reform of the criminal law to accommodate the practice of medically administered active voluntary euthanasia is appropriate – indeed necessary – in order to protect the interests of doctors and patients and to promote the principles of autonomy and self-determination. Whether or not one accepts her argument in its entirety, this detailed study is essential reading for anyone with an interest in the subject.

The book is divided into eight chapters. Chapters one and two present a detailed picture of the current legal situation with regard to euthanasia and assisted suicide, providing an essential backdrop for the author's later argument in favour of legal reform. In the first chapter, she challenges the 'problematic and unsatisfactory' legal distinction drawn between active and passive euthanasia – a distinction which has had a profound influence on current medical practice. She directs our attention to the criminal law's incoherence: a doctor who administers active euthanasia at a patient's request faces criminal liability for murder in all common law jurisdictions, whereas a doctor who, at the request of a patient, withholds or withdraws treatment with the same intention of causing the patient's death, will not be held criminally responsible. Thus, even though 'the object and end result of active and passive euthanasia are the same, the legal consequences are vastly different.'

In chapter two the focus shifts to assisted suicide, a statutory criminal offence in the jurisdictions under consideration (with the recent exception of the American state of Oregon). Otlowski criticises the clear line drawn consistently by the courts between 'suicide' and 'treatment refusal' as one which lacks 'substance' and leads to 'an undesirable distortion of legal principles' in order to avoid what, in many cases, are obvious conclusions. Even if persuaded by Otlowski's reasoning, many readers will find this aspect of her analysis contentious, not least because of the potential implications of discarding this distinction. There are certainly clear policy reasons for maintaining it.

In chapter three, Otlowski presents and analyses 'incontrovertible evidence' which demonstrates that euthanasia is being administered illegally by physicians with a certain amount of frequency, even though such physicians are rarely prosecuted. This divergence between theory and practice, she maintains, highlights the 'inadequacies' of the current legal situation and demonstrates the need for reform of the law.

Otlowski also challenges the widely-accepted doctrine of double effect. She argues that it is hypocritical for the law to condone the administration of life-shortening palliative drugs in order to relieve pain yet strictly prohibit the practice

© The Modern Law Review Limited 1999 (MLR 62:1, January) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

of active euthanasia. While in certain circumstances the distinction between them becomes blurred, the two practices implicate wholly different ethical, legal, and social policy concerns. Otlowski, however, fails to appreciate fundamental differences between the practices when she suggests that they are similar 'in terms of outcome and intention, and arguably, also morality.'

In the fourth chapter, Otlowski examines the main arguments for and against the legalisation of active voluntary euthanasia. The central argument for legalisation rests upon the principles of autonomy and self-determination. Advocates of euthanasia emphasise that respect for these principles is of paramount concern – just as an individual in a democratic society is free to manage the course of his or her life, so should an individual, it is suggested, have the freedom to control the time and manner of his or her death. The author touches upon a number of other familiar arguments in favour of legalisation (including 'the need to alleviate pain and suffering and to prevent cruelty' and the 'promotion of human dignity'), the cumulative effect of which, she concludes, is sufficient to constitute a *prima facie* case for reform of the current law.

The case against legalisation is separated into doctrinal arguments and pragmatic issues. The author dismisses the 'uncritical' doctrinal objections, although she delves deeply into the practical arguments. This is in keeping with what seems to be a growing tendency on the part of many commentators to eschew the moral questions and focus their attentions instead on the possible implementation and regulation of the practice of euthanasia. It is not surprising, then, that the so-called 'slippery slope' or 'wedge' argument has emerged as 'the most popular objection' to legalisation. The argument assumes some prominence in this book as well.

Slippery slope arguments are frequently misunderstood and improperly applied in the euthanasia debate. The vast majority of those writing on the subject mistakenly consider the slippery slope issue to be solely a question of 'regulation'. Otlowski, on the other hand, correctly recognises that the argument includes a 'logical' component as well. However, her analysis of this version of the argument is weak, and she gives it only cursory treatment. The argument asserts that once we permit voluntary euthanasia, we are logically committed also to permitting at least non-voluntary (and perhaps involuntary) euthanasia. The usual response to this argument is simply to call attention to an underlying distinction between the different kinds of euthanasia without offering anything further. Otlowski's reply is no exception:

> a very real and significant distinction exists between voluntary euthanasia on the one hand, and involuntary or even non-voluntary euthanasia on the other, and this distinction must be firmly borne in mind in evaluating the wedge argument. Provided that the individual's choice is treated as determinative, there is a sufficiently clear line to prevent the imposition of active euthanasia on non-consenting patients ... Recognition of patient autonomy and self-determination is thus a clear limiting principle against abuse.

However, is it not also true that respecting a patient's autonomy and self-determination, and acknowledging that a prolonged existence may be of no value to a patient, may provide some justification for *non-voluntary* euthanasia as well? Furthermore, given Otlowski's earlier acknowledgement that 'mercy' is a critical consideration in the case for legalisation of voluntary euthanasia, is she able to justify the requirement of a contemporaneous 'request' of a competent patient? Certainly in law, the incompetent do not lose their rights merely because of the fact of their incompetence. Arguably, a strong case exists for anyone who would 'benefit' from euthanasia, even in the absence of any request on the patient's part.

© The Modern Law Review Limited 1999

The practical version of the slippery slope argument is given more significance in the book. Otlowski places the burden upon opponents of legalisation 'to provide convincing evidence' that undesirable consequences are 'likely', although it is arguable that this is not where the burden ought to fall. Nonetheless, it is here, the author argues, that the wedge objection founders:

> there is simply no compelling evidence to suggest that the acceptance of active voluntary euthanasia on a strictly limited basis would be the 'thin edge of the wedge'. Indeed there is significant evidence to the contrary; since 1970, active voluntary euthanasia has been practised in the Netherlands, yet, contrary to the claims of some opponents, there is no convincing evidence to suggest that this has represented a step along the slippery slope, leading to involuntary euthanasia

This statement will seem remarkable to those who are familiar with extensive research documenting alarming instances of abuse in the Netherlands.

Otlowski also contends that euthanasia is currently being practised in a clandestine fashion, without any protective safeguards. She insists that legalisation will 'reduce the possibility of covert and improper conduct.' However, nothing in Otlowski's argument demonstrates convincingly that a law that is infringed with some degree of frequency *before* legalisation will not be infringed *after* it has been legalised. Indeed, it seems that her position actually lends support to the slippery slope argument.

In chapter five, Otlowski changes her focus to the factors which have given rise to the 'receptive climate for reform'. Her extensive treatment of contemporary developments, both in this chapter and in others, is one of the book's strengths. In chapter six, Otlowski outlines the various 'official' responses to the call for legal reform. She devotes considerable space to the developments in the Northern Territory of Australia. Although, at the time of the book's publication, this was the first jurisdiction in the world to legalise euthanasia and physician-assisted suicide, the legislation has since been overturned, thereby rendering significant portions of this chapter obsolete.

In the final chapter, Otlowski sets out the possible options for reform of the law. After considering a variety of models, she suggests that outright legalisation of active voluntary euthanasia in accordance with strict criteria would be the best approach. Her honesty is commendable, especially given that attention has shifted of late within the literature to the more fashionable topic of 'physician-assisted suicide'. It has been suggested by some commentators that a bright line may be drawn between the two practices, and that the legalisation of assisted suicide poses a lesser risk of abuse than does that of euthanasia. However, as Otlowski acknowledges, the more convincing argument seems to be that, from a practical viewpoint, the two practices are virtually interchangeable. And as 'there will always be a proportion of patients who are physically unable to commit suicide,' any attempt to limit legal accommodation to assisted suicide would be indefensible in principle.

Otlowski canvasses the possible safeguards to minimise the potential for abuse, and remains extraordinarily optimistic about the viability of a legislative framework based largely upon the Dutch model. The Netherlands experience, however, demonstrates that her optimism may be rather misguided, and calls into serious question the feasibility of regulation. Moreover, Otlowski is unable to overcome what may be the fundamental impediment to effective monitoring and regulation – namely, the inherently private nature of the doctor-patient relationship. One must seriously question whether *any* system of regulation can

© The Modern Law Review Limited 1999

offer protection against the unauthorised practice of euthanasia within a confidential and personal domain such as this.

While many may agree with the author's assessment of the current legal framework and its inherent difficulties, not all will be led to conclude, as Otlowski has, that legalisation of euthanasia is the correct response at this time. What she proposes in this book is a drastic and fundamental alteration of the criminal law and the practice of medicine. One can, therefore, predict that no reader, regardless of his or her position on euthanasia, will fail to be challenged by Otlowski's contentions. While the book's shortcomings undermine the cogency of her arguments, she unquestionably succeeds in advancing a crucial debate in a meaningful way.

*Wendy E. Hiscox**

Francis D. Rose (ed), **Failure of Contracts: Contractual, Restitutionary and Proprietary Consequences**, Oxford: Hart Publishing, 1997, xxvi + 294pp, pb £30.00.

The editor acknowledges that the papers in this book range over a number of different themes and that the book is not intended to form a comprehensive treatise dealing systematically with all matters concerning the failure of contracts. Yet the collection does deal with those aspects of failure of contracts which remain most controversial in English law. More importantly, the title of the book is somewhat misleading in that it does not reveal the existence of Hedley's essay which constitutes a major contribution to the debate about the current perception of restitutión in English legal theory. A fundamental point is in issue here. As is acknowledged in the comment on Hedley's essay, the English market for writing on the law of restitution is approaching 'saturation point'. My favourable review of this book, and any critical comments which I make, stem from my sharing most (but not all) of Hedley's views.

Ruxley Electronics and Construction Ltd v *Forsyth* [1996] AC 344 is an important decision and well worth the cogent analysis and criticism contained in O'Sullivan's essay. Given that it is O'Sullivan's belief that the decision of the House of Lords, however welcome, will make it virtually impossible to advise appropriate settlement terms, I would query the appropriateness of raising hypothetical restitutionary analysis whereby a builder saves costs in performing defectively. This situation did not, of course, arise in the case and, I suspect, would involve extremely difficult issues of evidence. I had not appreciated the distortion of the facts at the different stages of the case and share O'Sullivan's concern over this matter. Nolan, in his examination of a familiar line of controversial cases, pursues in greater depth a point which was incidental to O'Sullivan's essay. The subject of Nolan's essay is the availability and nature of remedies when a contract-breaker profits from his breach. Part of his conclusion is that the remedy of restitution for breach of contract, with one exception, is not compatible with current contractual doctrine. However, he points out, rightly in my opinion, that it does not follow that current contractual doctrine cannot accommodate a change to the law of remedies for breach.

* University of Cambridge

 © The Modern Law Review Limited 1999

Worthington's essay concerns the proprietary consequences of contract failure. Much of her discussion centres around the decision of the House of Lords in *Westdeutsche Landesbank Girozentrale* v *Islington London Borough Council* [1996] AC 669. In criticising the outcome of that case, she concludes that proprietary interests arise not only when contracting parties are obliged to perform their obligations but, also, when they are obliged to undo their dealings. I doubt her suggestion that this conclusion would not deliver unacceptable commercial consequences in insolvency cases, and although I find her propositions interesting, I cannot envisage future adoption by the judiciary as a likely occurrence.

Toube's essay, addressing the question of whether restitution under Article 85(2) of the EC Treaty can be achieved, is the first in the collection to introduce an international element. Toube is admittedly cautious in offering an exploratory introduction to a highly complex area. Her examination of the shift in English law from 'failure of consideration' to 'absence of consideration' is nevertheless of considerable importance. The title of Dannemann's essay is misleading since it suggests that he is examining restitution for termination of contract in German law. He in fact offers a comparative analysis with English law which reveals illuminating differences in outcome where a contract breaker claims restitution. The problems of an innocent party, or a party in breach seeking restitution of money paid or restitution for benefits in kind, are lucidly explained. Schanze's essay continues the comparative theme dealing with the topic of the failure of long-term contracts and the duty to negotiate. He expresses grave doubts about the impact of paternalism in this particular area of contract law, pointing out that *Williams* v *Roffey* [1991] 1 QB 1 allows an option for re-negotiation and does not prescribe any legal rule as to a duty to re-negotiate.

Failure of contract is peripheral to MacQueen's essay, which considers the general question of using the concept of unjust enrichment in contract law. MacQueen concludes that use of the concept would fatally undermine the fundamental justification for enforcing contracts. Hedley's essay takes this general argument further and has nothing to do with failure of contracts. It is an examination of the paradox involved in the decay of contract coinciding with the coming-together of restitution. Hedley is extremely critical of the theoretical structuring of the law of restitution which has occurred since the publication of Birks' *An Introduction to the Law of Restitution* in 1985. He acknowledges that his scepticism about the adoption of any theoretical structure is not the way to instant popularity, as it suggests that rather a lot of people, academics in particular, have been wasting their time. I believe, nonetheless, that his essay is worth considering by both contract and restitution lawyers. His general sketch of a model of contract law might well be recommended as a syllabus for contract law courses. It constitutes a de-construction of contract law. For example, he explains that the principle which states that 'contracting parties are absolutely bound by their agreements' is modified in cases of misrepresentation and non-disclosure, duress or undue influence, abuse of superior bargaining strength, lack of good faith, mistake, variation by the parties, frustration and relational contracts. His objective is to demonstrate that contract law no longer fits within a neat theoretical structure. Contract law teachers might do well to consider whether their approach is, in Hedley's terms, 'traditional', 'modernist' or 'post-modern'. Hedley contrasts the decay of contract theory with the unprecedented and frenetic pace with which restitutionary theory has developed in the English jurisdiction. The point of this comparison is to demonstrate, in contrast to contemporary perceptions, that restitution is highly miscellaneous and needs no all-embracing theory.

© The Modern Law Review Limited 1999

In conclusion, this is a collection of interesting and highly readable essays. It makes a significant contribution to a fascinating and ever-developing discussion concerning the evolution of law, both in practice and in theory.

*Margaret Halliwell**

Duncan Kennedy, **A Critique of Adjudication (fin de siècle)**, Cambridge, Massachusetts: Harvard University Press, 1997, iv + 424 pp, hb £29.95.

Duncan Kennedy's new book, *A Critique of Adjudication*, sets out to outline 'a theory of the political effects of the American social practice of organizing law making through distinct adjudicative and legislative institutions'. It 'is a work of general social theory written from a leftist and a modernist/postmodernist point of view' and 'represents adjudication and legislation as parts of political and cultural life, in the hope that the picture presented will be both convincing and unsettling' (p 1). The book skilfully achieves its intended objectives and provides us with a remarkable and original approach to the problems of jurisprudence. But it will still disappoint those who expected a major theoretical statement or a new 'canonical text' from one of the foremost critical legal theorists. Kennedy does not believe that such an abstract theory is possible at all and he is not offering one. Furthermore, Kennedy's argument follows in the footsteps of a wider philosophical tradition of critical, phenomenological or interpretative social and political thought. Hence, it does not need to return to and revamp the critical methodologies of the great practitioners of the genre, such as Marcuse, Horkheimer, Merleau-Ponty, Sartre, Foucault or Derrida. This omission may disappoint those – mainly analytical jurisprudes – who want books of legal theory to be as self-contained as possible, but it cannot be a valid criticism against this book. The critical tradition is there. It may not be exactly transparent or lucid or to everyone's taste, but anyone can dip into it, if they so wish.

Kennedy's book is a mixture of social theory, cultural studies, legal doctrine and philosophical reflection. The main theme is a theoretical account of legal reasoning in appellate adjudication. This is offered against the background of a careful reconstruction of recent American legal and political history and an (openly autobiographical) account of the development of the cls movement in response to this history. Kennedy's informal style is often suggestive and provocative, sometimes liberating and at times irritating, especially when it gets in the way of the argument. At one point early in the book, Kennedy begins a section the following way: 'This morning I woke up thinking about how to start this section and read the Sunday newspaper' (p 143). He then goes on to cite and discuss at length a report in the newspaper of a New York case regarding liability waivers at a university dental clinic. On the basis of that case he then proceeds to discuss complex issues of tort liability etc. It is certainly a style very few in legal academia will attempt to emulate.

The book's main project is to examine 'the place of adjudication in an ideologically divided society' (p 39). 'Ideology' is used to denote political projects and beliefs linked to group interests but with some claim to 'universal' rightness or practical validity. For Kennedy, 'both liberalism and conservatism are ideologies because these quite concrete positions in group conflicts are backed up by more or

* Faculty of Law, University of Manchester.

 © The Modern Law Review Limited 1999

less elaborate universalization projects, which allow advocates to claim that each of the more particular positions is an instance of correct application of general principles' (p 47). The argument begins by describing the ways in which constitutional adjudication plays a central political role in contemporary United States (chapters 4 and 5) and then, in the central part of the book, proceeds to explore in detail how 'ideology' shapes legal argument and makes the identification and interpretation of rules possible (chapters 6 to 8). Two other parts deal with the ideological consequences of adjudication as a social practice (chapters 9 to 11) and the ideological significance of rights in American legal consciousness (chapters 12 and 13). The last chapter, where, under the title 'Landscapes along the Highway of Infinite Regress', Kennedy muses on his own 'modernist/postmodernist project', is one of the most intense and suggestive parts of the book.

It is impossible to summarise here the various aspects of this impressive work, much less comment on them. Instead, I shall only focus on the classic question of the separation of law and morals. It would be a serious mistake to assume that Kennedy's argument is the simplistic one that law is politics by a different name. The point is clearly recognised that judicial fidelity to the legal materials does make a difference and does constrain the judge in a number of ways. Kennedy tells us, for example, that 'judges are, to a significant extent, practically "bound" by law and often, often, often declare and apply rules that they would never vote for if they were legislators' (p 275). However, because of the open-ended nature of the legal materials it is always possible or necessary for the judge to make choices on the basis of evaluations of an 'ideological' nature. Kennedy writes that 'legal materials and legal reasoning are sufficiently plastic that they can offer an acceptable post hoc rationalization of whatever result the judge favors, and judges are habitual rationalizers' (p 159). The point is that even before one starts evaluating conflicting arguments for a particular outcome, one has already done 'ideologically oriented legal work' even in the process of recognising and setting up the 'field' where the legal question supposedly lies. 'Ideological' or general evaluative questions arise at every turn in the process of legal argument and rule-application.

Although Kennedy would probably deny this, at the core of his argument lies a simple insight about the general role of rules in practical reasoning, an insight shared by critical theory, hermeneutics and Dworkinian legal theory. The point is that rules do not determine their own application in concrete instances, but always depend on a complex set of presuppositions, interpretative work and other meta-rules. This pragmatic view of rules, explored by Kant, Hegel and Wittgenstein among others, must be at the root of Kennedy's account of rule-following and determinacy: 'The question, "does this question of law have a determinate answer?" is therefore meaningless if it is a question of law, rather than a question about the interaction between a particular, situated historical actor and this particular question of law situated in this particular field' (p 170). Kennedy's project may be seen as part of broader philosophical critique of rules and rule-based practical reasoning. It can thus be read alongside recent work on legal argumentation by authors such as Stanley Fish, Sanford Levinson or Jack Balkin.

But Kennedy is very keen to avoid such comparisons and follows his own version of contextualism and particularism. He rejects any ambitious attempt to arrive at a programmatic or abstract critique, which he calls 'Golden Bowlism', the belief that 'a given formal discourse will shatter if only one can find *the* crucial pressure point'. In fact, the situation is more complex: 'There is no single principle

© The Modern Law Review Limited 1999

or metaprinciple, at least as far as I can see, that links together all the different kinds of argument that we want to call "formalist", except that they all fall to internal critique. And there is no general principle, as far as I can see, that links together all the different strategies of the internal critique into a single supercritical unity. So there is no guarantee, as far as I can see, either that there will or won't be a critique available for every assertion of legal necessity' (p 277). This is why it is necessary, for Kennedy, that the critic not just put forward the critique of rules in general, but also that he or she pursues it diligently in the various areas of legal scholarship.

Now this is an original and interesting position and has been elegantly argued. But is there anything particularly left wing or radical about it? Does it follow from, or lead towards a radical plan of social transformation? Kennedy, departing here from Unger's critical project, does not think so. His leftism does not proceed from any deeper theoretical commitments. It is presented as a choice which the author clearly does not wish to defend or discuss in grand theoretical terms. Of course, the question then arises, why should anyone else join him in his politics, if they did not have the same instinctive identification with the left in the first place? In other words, is there anything to leftist social and political theory, other than an instinctive sympathy for left wing politics? This question is not answered by Kennedy's book, which is by no means a political manifesto.

This morning, just before I set out to finish this review (and for entirely unrelated reasons), I visited Karl Marx's grave, a few minutes' walk from my flat in London. It is an ugly memorial, crowned by a huge bust of the great man. Below the names of Marx, his wife and other relatives who are buried there, there is a large inscription with his dictum that philosophers should not just set out to know the world, but should also change it. Kennedy has taken this to heart, and constantly reflects on possible avenues of action for 'leftist, modernist-postmodernist' academics. His conclusion is intriguing: it ends with a defence of 'politicization and passivity', by which he means 'trying to set up a political identity to the left of liberal bad faith, without being or seeming to be a wrecker' (p 374). But change the world to what? Marx gave his followers a false promise of an ultimate liberation from bourgeois misery and injustice through the total transformation of society. Fortunately, very few of today's radicals would claim that, in the light of some well known attempts, such a total transformation of ordinary life through politics is plausible or desirable. Kennedy would certainly not be one of them. So what is left? Kennedy's main argument may well be right that liberal judges are constantly in 'bad faith' and 'denial' over the place of ideology in adjudication. But what exactly is wrong with liberalism?

*Pavlos Eleftheriadis**

* London School of Economics and Political Science.

© The Modern Law Review Limited 1999

Information for Contributors

Contributions to and correspondence concerning the various sections of the Review should be sent to the relevant Editor:

Articles: Hugh Collins, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Legislation and Reports: R.C. Simpson, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Case Notes: Roger Brownsword, Faculty of Law, The University, Crookesmoor Building, Conduit Road, Sheffield S10 1FL.

Books for Review, Review Articles and Reviews: Neil Duxbury, Faculty of Law, University of Manchester, Oxford Road, Manchester M13 9PL.

General Correspondence: Tim Murphy, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Section Editors will be happy to discuss drafts and proposed contributions with contributors. All contributions should be typed, double spaced and include an estimate of the number of words, including footnotes. Contributions should use footnotes which make cross-references within the text as sparingly as possible. Contributions must conform to the Review's house style; a style sheet will be provided by the Production Editor on request.

Articles do not usually exceed 12,000 words (excluding footnotes). Longer or shorter articles can be considered.

Subscription Information 1999

The Modern Law Review is published six times a year. Subscription rates are as follows:

Annual subscription for unbound parts only

UK and Europe	£73.00
Rest of the world	£75.00
North America	US$124.00*
UK and European law teachers and law students	£21.00
Commonwealth law teachers and law students	£26.00/US$43.00*

Annual subscriptions for Bound Volume Service: unbound parts plus a bound volume

UK and Europe	£121.00
Rest of the world	£128.00
North America	US$211.00*
UK and European law teachers and students	£59.00
Commonwealth law teachers and law students	£62.00/US$102.00*

Back Issues

Single issues of the unbound parts of the last three volumes are available from the publishers at the current price for unbound parts:

	£15.00 (£15.00 for ROW, US$25.00 in North America)*
Single bound volumes	£61.00 (£63.00 for ROW, US$104.00 in North America)*
Complete sets of bound volumes	£2071.00 (US$3417 00 in North America)*

*Canadian customers/residents please add 7% GST.

Orders, sample copy requests, claims and other subscription enquiries should be addressed to: Blackwell Publishers Journals, PO Box 805, 108 Cowley Road, Oxford OX4 1FH, UK. All subscriptions are supplied on a calendar year basis (January to December).

US Mailing

Periodicals postage paid at Rahway, New Jersey.
Postmaster: send address corrections to *The Modern Law Review*, c/o Mercury Airfreight International Ltd Inc., 265 Blair Road, Avenel, NJ 07001, USA (US mailing agent).

Microform

The journal is available on microfilm (16 mm or 35 mm) or 105 mm microfiche from the Serials Acquisitions Department, University Microfilms Inc, 300 North Zeeb Road, Ann Arbor, MI 48106, USA.

CONTENTS

ARTICLES

Fiduciary Responsibilities — Andrew Kull — 1

Freight and Security in — Lois Mulcahy
Hostages the Peppercorn — Scott Turow
Power on the Second — [illegible]
Penalty Section — [illegible] — 11

Corporate Executives and the Modern
Shareholder Ownership — Randall S. Thomas — 33

Uncertainty, Determinacy — Ronald M. Noyes, Randall
and False Inference — [illegible] — 75

LEGISLATIVE

The Human Rights Act in Parliamentary
Democracies — Lisa Forman — 79

Law and Landscape: The Legal Construction
and Preservation of Wilderness — Eric Freyfogle — 105

CASES

Court Application for Injunction in Favor
The Unborn Child — Alison Mitchell Bass — 113

Personal Injuries and Death: Collateral Benefits
Collateral Recovery — Exclusive Remedy — 125

Director's Duties and Liabilities: Contracts
Between Companies Law
Fiduciary Duty of Loyalty — Fiduciary — 137

REVIEW ARTICLE

Equity's Deadly Bane: Its Future? — Victor Brudney — 155

REVIEWS

Corpus Juris Volumes: Parliament and the Common Law — 161
Personal Liability in Contract: Commercial
Reputations and Property Consequences — 163
Remedies: A Comparative Analysis of Damages Remedies — 165

0026-7961(199901)62 1:1-4

VOL. 62. NO. 2 MARCH 1999

THE MODERN LAW REVIEW

PUBLISHED BY THE MODERN LAW REVIEW LIMITED

THE MODERN LAW REVIEW

Volume 62

EDITORIAL BOARD

J.S. Anderson
M.R. Chesterman
Lord Dahrendorf of Clare Market
Gerald Dworkin
J.M. Evans
Anthony Giddens
Cyril Glasser
J.A.G. Griffith
Judge F Mancini
Harvey McGregor
T.M. Partington
I.G. Patel
Lord Wedderburn of Charlton, *Editor Emeritus*
Michael Zander

EDITORIAL COMMITTEE

Julia Black, *Production Editor*
Roger Brownsword, *Case Note Editor*
Hugh Collins, *Articles Editor*
W.R. Cornish
Neil Duxbury, *Reviews Editor*
Judith Freedman
Carol Harlow
Nicola Lacey
Jenny Levin
Martin Loughlin
W.T. Murphy, *General Editor*
R.W. Rawlings
S.A. Roberts
Colin Scott, *Secretary*
R.C. Simpson, *Legislation and Reports Editor*
F.G. Snyder
Erika Szyszczak

Published for The Modern Law Review Limited by Blackwell Publishers

© The Modern Law Review Limited 1999

ISSN 0026-7961

The Modern Law Review Ltd is incorporated in England as a company limited by guarantee under the Companies Act (Reg No 325 282) and is registered as a charity (Reg No 264072)

The Modern Law Review is published six times a year, in January, March, May, July, September and November by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main St, Malden, MA 02148, USA.

Information for contributors, subscription information and details about microfilm and advertising appear on the inside back cover.

The Modern Law Review website is at *http://www.lse.ac.uk/depts/law/mlr.htm*

Printed and bound in Great Britain by The Alden Press, Oxford
This journal is printed on acid-free paper.

COPYRIGHT: All rights reserved. Apart from fair dealing for the purposes of research or private study, or criticism or review, as permitted under the UK Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing of the Publisher, or in accordance with the terms of photocopying licences issued by organisations authorised by the Publisher to administer reprographic rights, such as the Copyright Licensing Agency and the Copyright Clearance Center.

*** FREE COURSEPACK COPYING IN UK TEACHING INSTITUTIONS DURING 1999***

For a trial period only (1 January to 31 December 1999) CLA copyright fees and permissions for the inclusion of articles, notes and reviews from any volume of Modern Law Review in study packs for use by students in UK teaching institutions have been waived. Such copies may be made free of charge during this period provided the institution does not profit from such activites

CONTENTS

ARTICLES

Human Rights and the House of Lords 159
Rt Hon Lord Hoffmann

Standards of Proof in Civil Litigation 167
Mike Redmayne

Theorising Utility Regulation 196
Tony Prosser

Reconsidering Disgorgement for Wrongs 218
Sarah Worthington

LEGISLATION

Unfinished Business: The Scotland Act 1998 241
Noreen Burrows

Detecting, Prosecuting and Punishing Benefit Fraud: The Social Security Administration (Fraud) Act 1997 261
Gráinne McKeever

CASES

Profits from Wrongdoing: Private and Public Law Perspectives 271
Margaret Halliwell

Contributory Negligence Applied to Economic Loss: *Platform Home Loans* and *Fancy and Jackson* 281
Tony Dugdale

'Hearsay-fiddles' in the House of Lords 290
John Hartshorne and Andrew L-T. Choo

REVIEW ARTICLE

Secrets of the Political Constitution 298
Ian Leigh

REVIEWS

McCoubrey: The Obligations to Obey in Legal Theory 310
Halpin: Rights and Law: Analysis and Theory 314
Harrison: Good Faith in Sales 316
Weir: Economic Torts 320
Prosser: Law and the Regulators 323
Goodrich (ed): Law and the Unconscious: A Legendre Reader 326
Sheldon: Beyond Control: Medical Power and Abortion Law 328
Amato: Antitrust and the Bounds of Power: The Dilemma of Liberal Democracy in the History of the Market 330
Zeisel and Kaye: Prove It With Figures: Empirical Methods in Law and Litigation 331
Cavadino: The Law of Gravity 332

THE
MODERN LAW REVIEW

Volume 62 March 1999 No 2

Human Rights and the House of Lords

*Rt Hon Lord Hoffmann**

The coming of the millennium is traditionally associated with a belief, going back at least to the Books of Daniel and Revelations, that the Kingdom of Christ or some other just ruler will be established, there will be peace and plenty upon earth, and those who have been unjustly treated will receive their rightful rewards. By way of encouragement of this belief, the Government have made the theatrical gesture of timing the coming into force of the Human Rights Act 1998 to coincide with the beginning of the new millennium. The press has cast the Law Lords in the improbable role of being the just rulers whose reign the Act will establish, the Platonic guardians, as Judge Learned Hand called them, who will put the nation to rights from the reformed House of Lords or, perhaps more probably, from a new Supreme Court established in some office block in the Strand. In anticipation of this accession of power, there have been analyses of the views and predilections of individual Law Lords, and they have all been issued with convenient party badges saying liberal, conservative, middle of the road and so forth.

I am bound to say that I do not see it that way at all. I suppose the image which some people have of the future of our supreme court, whatever it may be called, after the coming into force of the Human Rights Act, is based upon the model of the Supreme Court of the United States and to some extent the Bundesverfassungsgericht or Federal Constitutional Court of Germany, these being the courts engaged in the interpretation of basic rights which are closest to hand. And there is no doubt that both those courts have given decisions, based upon constitutional provisions, with a high political content. One has only to mention the American decisions on race segregation and abortion and the German decisions on the Maastrict treaty and the use of the German armed forces abroad. So it is assumed that once the House of Lords has got its feet under the constitutional table, it will find itself having to resolve similar highly charged issues.

I think that this assessment ignores some very considerable differences between this country on the one hand and the United States and Germany on the other. I put aside at once the technical distinction that the American and German courts can declare acts of the legislature to be unconstitutional and void whereas the Human Rights Act, in deference to the theory of the sovereignty of Parliament, allows the UK courts only to declare that legislation is incompatible with the European

* A lecture delivered to the Common Law Bar Association on 25 November 1998.

© The Modern Law Review Limited 1999 (MLR 62:2, March). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

Convention. I call this a technical distinction because I think that if the courts make a declaration of incompatibility, the political pressure upon the government and Parliament to bring the law into line will be hard to resist. The real differences seem to me to reside in the history, cultures, and political structures of the three countries.

If one examines the cases in which the United States Supreme Court has most obviously intervened in the political field, it has been in those areas in which it was widely recognised that change was needed but the checks and balances of the country's political system paralysed the ability of both the executive and legislative branches of government to carry it through. The most obvious example is the great case of *Brown* v *Board of Education*[1] in which the Supreme Court outlawed segregation in schools. Segregation in the southern states was an issue on which, because of the alliance of expediency between racialist Democrats in the south and liberal Democrats in the north, there had been no movement since the Civil War. The social pressures on segregation had been brought to bursting point by the equality which black soldiers had experienced in the United States Army during the war. But the blacks were disenfranchised in the southern States and the national parties, courting southern white votes, were unable to act. The Supreme Court had to intervene in an attempt to stave off civil unrest. Their intervention was not without a constitutional price; federal judges found themselves running schools and organising the transportation of children on buses. These are not activities which, in an ideal society, judges should have to undertake. In retrospect however, no one would deny that the decision of the Supreme Court was necessary and right.

Similar inaction on the part of the legislative branch of government explains judicial intervention in other areas. The various decisions on police powers in the 1960s which reached their high point in *Miranda* v *Arizona*,[2] which held that statements by an accused person were inadmissible unless he had been cautioned and told of his right to a lawyer, represent an attempt to deal with a serious problem of widespread abuse of power by the police which politicians running for election on platforms promising to be tough on crime were unwilling to tackle. Likewise, the intervention of the German Constitutional Court into what some would regard as areas of foreign policy such as the Maastricht Treaty and the use of the German Army in former Yugoslavia were widely seen as being the result of the lack of a political will to face these questions.

The position in the United Kingdom is, I think, rather different. Parliament has had no difficulty in legislating in many areas of human rights. Starting in the 1960s, the legislation decriminalising suicide, abolishing the death penalty, against racial discrimination, regulating abortion, against sex discrimination, regulating the conduct of the police in the investigation of crime, and so on, has meant that a good deal of the human rights law which, in the United States has been constructed by the courts on the basis of the very general statements of principle in the Bill of Rights, has here been the subject of detailed legislation by Parliament. In most of these areas there is now a general political consensus on the subject. I say political consensus because that is not necessarily the same as a consensus among the people as a whole. It is, for example, well known that a considerable majority of the public would favour the reintroduction of the death penalty. But the politicians, taking a traditional Burkean view of their role as representatives and not delegates, have ensured and are likely to continue to ensure that this will not happen. They are

1 347 US 483 (1954)
2 384 US 436 (1966).

© The Modern Law Review Limited 1999

able to do so because of the strength and discipline of the party system in this country, which means that most voters are forced to buy the complete packages of one or other of the three major parties and politicians gain no advantage by advocating views which may have popular support but are unacceptable to the leadership of their parties. The result is that no one today can cast an effective vote in favour of the reintroduction of capital punishment. The English political system is thus in a sense aristocratic. In using this term I do not refer to the hereditary peers, but rather to the fact that in England politicians as a class are not directly accountable to the people. They are able, by virtue of the party system, to exercise an independent judgment.

This political consensus over a wide area of human rights does not exist in the United States, where the looseness of the party system allows more scope for anti-libertarian populism. That in turn produces a situation which requires the intervention of the courts, because not only are politicians often reluctant to deal with human rights abuses, but they are also inclined to pass populist measures which plainly contravene constitutional rights, leaving it to the courts to incur the odium of thwarting the will of the people.

The result of these various political strands is that judicial intervention is an accepted feature of American life, resulting in a public awareness of the importance of the Supreme Court and the views of the individual justices who are called upon to decide these issues. There is a similar German awareness of the role of the constitutional court, although the collegiate nature of that body meant, at least until recently, that there was not so much concern with the views of individual members. But this is most certainly not the case in the United Kingdom, where people expect such issues to be decided by Parliament, as they have been in the past, and Parliament, for the reasons I have mentioned, has on the whole been able to show a great sense of responsibility in dealing with them. Until very recently, the identities, let alone the views of the Law Lords were completely unknown to almost everyone.

I must admit that I like it that way and I feel particularly comfortable with the fact that, for example, the law on abortion has been made by the elected representatives of the people rather than being deduced from some very general statements in the Bill of Rights. I do not relish the role of a Platonic guardian and I am pleased to live in a society that does not thrust it upon me. Having said that, I do not want to sound too complacent; there are areas in which politicians are reluctant to enter, such as euthanasia, and other areas in which they have occasionally been perhaps too eager to enter, such as the rights of accused persons, prisoners and immigrants, in which there is plainly a role for the courts in the protection of human rights. On the whole, therefore, I would applaud the Human Rights Act and I think that there are areas of the law in which it will make a difference. But I do think that its potential impact has been greatly exaggerated.

I particularly hope that the passing of the Act will not in any way diminish the sense of responsibility which Parliament has always shown in passing legislation affecting human rights. It is a remarkable fact that during the period of more than two centuries in which human rights in the United States have been protected from Congress and the States by the enforcement by the courts of the articles of the Bill of Rights, Parliament has transformed this country into a democracy, in the sense not only of a country governed by a majority but also in the equally important sense of a country which respects the rights of individuals and minorities. For the whole of that period we have entrusted our most fundamental liberties to the will of a sovereign Parliament and, taken all in all, Parliament has not betrayed that trust.

© The Modern Law Review Limited 1999

It would be most regrettable if the knowledge that the courts were the ultimate arbiters of conformity with human rights resulted in Parliament being more ready to pass populist measures which infringed them. For my part, I attach great importance to the provision in section 19 of the Act, which requires a Minister of the Crown in charge of a Bill in either House to make a statement before the Second Reading that in his opinion the provisions of the Bill are compatible with Convention rights. Speaking for myself, I would take such a statement extremely seriously and would be very reluctant to decide that such a measure was not in fact compatible. And I think it is important that Ministers of the Crown should know that such statements will not be regarded as mere formalities but need to be made with careful thought and on responsible advice, because they will carry great weight with the courts.

Let me now move on to my second theme, which is the relationship, after the coming into force of the Act, between the courts of the United Kingdom and the European Court of Human Rights in Strasbourg. The only relevant provision in the Act is section 2(1)(a) which says that a court determining a question which has arisen in connection with a Convention right must take into account any judgment of the European Court of Human Rights. It must be taken into account, but is not binding. On the other hand, the right of individual petition to the court in Strasbourg is not being abolished, and it seems that the United Kingdom will continue to be under an obligation in international law to bring its domestic law into line with the Convention as interpreted by that Court. Is this a sensible state of affairs? We have had a very recent example of a decision of the Strasbourg court giving an interpretation to the Convention which, I venture to suggest, it is inconceivable that any domestic court in this country would have adopted. I want to use this decision as an example of the potential conflict between our legal system under the new regime and the jurisprudence emanating from Strasbourg.

On 28 October 1998, judgment was given in the case of *Osman* v *United Kingdom*.[3] Mrs Osman's husband had been shot and killed by an insane teacher who had formed an obsessional attachment to their young son at school. She sued the police for damages, alleging that they had been warned that the man was a danger to her family but had not taken adequate steps to protect them. The Court of Appeal struck out the action on the ground that the police could not be made liable in negligence for failing to take action in the investigation or suppression of crime.[4] The decision was based upon the Yorkshire Ripper case, *Hill* v *Chief Constable of West Yorkshire*,[5] in which the House of Lords held that the imposition of a duty of care in such circumstances would be contrary to public policy. The maintenance of police efficiency was better secured by other methods than having the question of whether they had acted reasonably in a given case expensively investigated in civil proceedings at the instance of a private litigant, with the possibility of compensation having to be paid out of the police budget. The prospect of such an investigation and the payment of compensation might in fact be detrimental to good policing, since it might make the police defensively unwilling to take risks. Furthermore, the efficiency with which the police handled a particular investigation often depended upon the share of their resources which was devoted to that kind of investigation and questions about the allocation of resources by public authorities such as the police were not suitable for determination by judges.

3 http://www.dhcour coe.tr/hudoc.
4 *Osman* v *Ferguson* (1993) 4 All ER 344.
5 [1989] AC 53.

© The Modern Law Review Limited 1999

The result of this decision was that as a matter of English substantive law, persons who suffer criminal damage have no claim to compensation solely on the ground that it could have been prevented by more efficient policing. This decision is in line with a number of recent cases on the failure of public authorities to deliver services which could have prevented loss, such as *X (Minors)* v *Bedfordshire County Council*[6] on social services, *Stovin* v *Wise*[7] on highway improvements, *Capital and Counties plc* v *Hampshire County Council*[8] on fire services, *O'Rourke* v *Camden LBC*[9] on housing the homeless, and *Murphy* v *Brentwood LBC*[10] on building inspection services. The theme which runs through these cases is that the fact that public services are provided at the public expense to confer benefits or protection on members of the public does not mean that a person who fails to receive those benefits or protection will be entitled to sue for compensation on the ground that the authority acted negligently in failing to provide them. The social justification for such a rule is that, on the one hand, the person who has failed to receive the benefit is no worse off than if it had not been provided in the first place, and on the other hand, the budgetary and efficiency grounds discussed in *Hill* v *Chief Constable of West Yorkshire*.

Not everyone would agree with such a rule. Some might think that it shows a somewhat niggardly attitude to people who have suffered loss because they had the misfortune not to receive some public benefit which they were reasonably entitled to expect, and that the allocation of funds for public services should budget not merely for providing the services but also for compensating those who did not receive them. It might also be said that the law is inconsistent in excluding liability in such cases but making public authorities liable when they enter into relationships which have long been recognised as giving rise to a duty of care in private law, such as between a doctor or hospital and a patient. For my part, I see no inconsistency, but I recognise that the merits of the rule are open to political debate, the question being essentially one about the obligations of the welfare state.

Mrs Osman, however, petitioned the European Court of Human Rights on the ground that the rejection of her claim was a denial of her fundamental human right under Article 6.1 of the Convention, which provides that 'in the determination of his civil rights and obligations, everyone is entitled to a hearing by a tribunal'. One might be forgiven for thinking that Mrs Osman's rights had been determined, rightly or wrongly, by a tribunal. She had been before the Court of Appeal and they had decided that even if all the allegations in her statement of claim were proved at the trial to be true, she would not be entitled under English domestic law to compensation from the police. But the Court, consisting of 17 judges including one from the United Kingdom, decided unanimously that because her action had not been allowed to proceed to trial, she had not had a hearing. What, you may ask, would be the point of a hearing which, under English law, was bound to end in the claim being dismissed? The answer, according to the European Court, was that it should not have been treated as bound to be dismissed. There should have been the possibility that on the facts alleged, Mrs Osman would win. A rule that in no circumstances should a person be able to claim compensation on the ground that the police failed to protect him from criminal injury was not proportionate to the

6 [1996] 3 All ER 353
7 [1996] 3 All ER 801.
8 [1997] 2 All ER 865.
9 [1997] 3 All ER 23.
10 [1991] 1 AC 398.

© The Modern Law Review Limited 1999

public policy grounds advanced in its support. In other words, English domestic law failed to provide compensation out of public funds in cases in which the Strasbourg court thought it should do so.

I am not sure, on a reading of the judgment of the court and the concurring judgment of the British judge, whether they had persuaded themselves that they were really dealing with the right to a hearing rather than the merits of the substantive tort law under which the Court of Appeal had held that Mrs Osman was bound to lose. In my view, there is no disguising the fact that the case was about the latter.

I am bound to say that this decision fills me with apprehension. Under the cover of an Article which says that everyone is entitled to have his civil rights and obligations determined by a tribunal, the European Court of Human Rights is taking upon itself to decide what the content of those civil rights should be. In so doing, it is challenging the autonomy of the courts and indeed the Parliament of the United Kingdom to deal with what are essentially social welfare questions involving budgetary limits and efficient public administration. I say the Parliament of the United Kingdom because it must follow from the decision of the Strasbourg court that even if the rule in *Hill* v *Chief Constable of West Yorkshire* had formed part of the Police Act 1996, it would have been held to contravene the right in Article 6.1 to a hearing before a tribunal. I understand that a petition by the unsuccessful plaintiffs in *X (Minors)* v *Bedfordshire County Council* is already on its way and no doubt there are other cases in which the same general principle will be challenged. The whole English jurisprudence on the liability of public authorities for failure to deliver public services is open to attack on the grounds that it violates the right to a hearing before a tribunal. And although the Strasbourg court appears to contemplate the possibility that after a hearing on the facts, the court might still come to the conclusion that on grounds of public policy the claim should be dismissed, it ignores one of the principal reasons for the present doctrine, which is to avoid a trial altogether, to avoid the waste of public resources involved in a judicial investigation, usually on legal aid, as to whether the public authority should reasonably have provided the benefit or not.

It may be that the court did not understand the rather formulaic reasoning by which English courts say that no duty of care exists when they mean that the case is one in which the law does not recognise a right to compensation. It may be that they did not understand the principle by which an action is struck out without going to trial if proof of all the facts alleged would not sustain a cause of action. Either way, the case serves to reinforce the doubts I have had for a long time about the suitability, at least for this country, of having questions of human rights determined by an international tribunal made up of judges from many countries.

I would not like anyone to think that this view reflected a vulgar Euroscepticism. When it comes to questions of a common currency and a large number of other economic and social issues, the advantages and disadvantages of a common European position are matters for pragmatic decision on which one has to weigh up the evidence as best one can. I do not regard national sovereignty on every issue as either possible or desirable. But the international adjudication of questions of human rights goes much deeper because it raises an ancient question about the universality of human values. It brings into focus the conflict between the universalist philosophy of the French enlightenment, who thought that French concepts of *liberté, égalité, fraternité* were derived from our common humanity and could therefore be applied world-wide, and pluralist views such as those of Goethe's friend, Herder and, in our own time the late Sir Isaiah Berlin, who said

© The Modern Law Review Limited 1999

that each community had its own set of values, different but not better or worse than those of other communities.

Voltaire said that morality was the same in all civilised nations. This is a half truth; of course we share a common humanity and there are some forms of behaviour such as torture which we all either reject or are unwilling to acknowledge. But even this is not inherent in civilisation: in the law of ancient Rome, the evidence of slaves was always taken under torture and none of the great Roman lawyers who developed the elaborate and sophisticated system, which forms the basis of law through so much of the world, seems to have thought this inhuman. Of course, I applaud the patient efforts of the human rights movement since the Second World War to promote the acceptance of basic human rights throughout the world. I am well aware that there are countries which deny their citizens basic human rights and which claim to be justified on grounds of cultural diversity, when the true reasons are the power and greed of their rulers. Nevertheless, I say that Voltaire's remark was only half true and that in a confident democracy such as the United Kingdom the other half is important. We do have our own hierarchy of moral values, our own culturally-determined sense of what is fair and unfair, and I think it would be wrong to submerge this under a pan-European jurisprudence of human rights.

The problem about the hierarchy of rights is not the conflict between good and evil but the conflict between good and good. Free speech is a good thing; justice is a good thing, but there are cases in which free speech and justice come into conflict with each other. For example, the law that preserves the anonymity of rape victims is an infringement of the freedom of the press, but it assists justice by encouraging women to make complaints against rapists. How then are these two desirable objectives – free speech and justice – to be reconciled with each other? There is no right answer to that question; any choice involves some degree of sacrifice. But in my view, the specific answers, the degree to which weight is given to one desirable objectives rather than another, will be culturally determined. Different communities will, through their legislatures and judges, adopt the answers which they think suit them. So the Supreme Court of the United States has ruled a State law which gave anonymity to rape victims to be unconstitutional as infringing the First Amendment right to freedom of the press, whereas such laws exist in the United Kingdom and other countries.[11] The Supreme Court of the United States attaches what other communities may regard as an exaggerated value to the First Amendment and no doubt they think we do not respect it enough. But the difference in the answers given to the perennial conflict between goals which everyone accepts to be good shows how impractical it would be to subject the United States and the United Kingdom to a common court of human rights. If this is true of two countries so closely bound together in culture, law and history, how much more true must it be of the disparate collection of states, some old democracies, some former police states, which belong to the Council of Europe.

Of course it is true that to some extent the Strasbourg court acknowledges the fact that often there is no right answer by allowing what it calls a 'margin of appreciation' to the legislature or courts of a member State. Within limits, they are allowed to differ. And, as I have said, I accept that there is an irreducible minimum of human rights which must be universally true. But most of the jurisprudence which comes out of Strasbourg is not about the irreducible minimum. These questions tend to come into play at a point when civil society itself is called into

11 *Cox Broadcasting Corporation* v *Cohn*, 420 US 469 (1975); Sexual Offences (Amendment) Act 1992.

© The Modern Law Review Limited 1999

question; they are far from the normal currency of dispute over competing values in a democracy like the United Kingdom. The *Osman* case, dealing with the substantive civil law right to financial compensation for not receiving the benefit of a social service, is as far as one can imagine from basic human rights. And I have taken it only as a very recent example: it is by no means unusual. It is often said that the tendency of every court is to increase its jurisdiction and the Strasbourg court is no exception. So far as the margin of appreciation accommodates national choices, the jurisdiction of the European court is unnecessary; so far as it does not, it is undesirable.

I would not deny that there have been cases in the history of the Strasbourg court in which the United Kingdom legislature and government have done things which a United Kingdom court, construing the Convention, would also have said amounted to an infringement of human rights. In those cases, in which the courts of the United Kingdom were unable to apply the Convention themselves, the Strasbourg court has performed a service by requiring the UK government to amend the law. But with the coming into force of the Human Rights Act, it will no longer be necessary to rely upon Strasbourg for this purpose. I realise that in the present political climate, it would be very difficult for the government to withdraw from the Strasbourg court. When we joined, indeed, took the lead in the negotiation of the European Convention, it was not because we thought it would affect our own law, but because we thought it right to set an example for others and to help to ensure that all the Member States respected those basic human rights which were not culturally determined but reflected our common humanity. There is still a great need throughout the world for the enforcement of such minimal human rights. But the jurisprudence of the Strasbourg court does create a dilemma because it seems to me to have passed far beyond its original modest ambitions and is seeking to impose a Voltairean uniformity of values upon all the member States. This I hope we shall resist. The White Paper which introduced the Human Rights Bill was called *Rights Brought Home*.[12] Now that they have been brought home, I hope that we shall be able to keep them here.

12 Cm 3782 (1997).

© The Modern Law Review Limited 1999

Standards of Proof in Civil Litigation

Mike Redmayne*

It is well known that the standard of proof in a civil case is proof on the balance of probabilities, and that this means that the party bearing the burden of proof must prove that her case is more probable than not. Indeed, the civil standard of proof appears to be one of the simplest concepts in the law of evidence, requiring little explanation or illustration. But scratch the surface of this most basic of evidentiary notions and an altogether more complex picture is revealed: the case law provides a range of conflicting interpretations of what the civil standard of proof requires in different contexts. When an area of the law is this confused, one starts to suspect that the problem lies in more than a failure by the appellate courts to resolve conflicting authorities and to lay down clear guidance (though this has certainly added to the difficulties in this area); one is drawn instead to the conclusion that the confusion lies at a deeper, conceptual level and that it is driven by the lack of a clear understanding of the basic building blocks of forensic proof.

In the first section of this article, I describe a simple set of conceptual tools for analysing the civil standard of proof. With this foundation in place, I then survey and evaluate the various approaches to the civil standard of proof that have been developed in the case law. The final section of the article is more ambitious: I discuss three situations in which judges have held that variations on the basic civil standard of proof are required and assess whether or not some variation is justified. The analysis in this final section may prove controversial, but whether or not readers agree with it, I hope that the preceding sections will, at least, have provided some common ground on which debates about these matters can take place.

Probability and utility: two components of forensic proof

In a typical civil case, two parties dispute whether one of them has broken some rule.[1] If the court decides that one of them has broken the rule, damages will be awarded to the other party. The court will receive and evaluate evidence because, in order to determine whether the rule has been broken, it must decide what happened. The court can never be completely certain about what happened in the past, therefore this factual determination need only be made to a degree of probability. The concept of probability is the first important component of proof. It

* Department of Law, Brunel University.

I am grateful to Andrew Choo, Neil Duxbury, Richard Friedman and Felicity Kaganas for comments on an earlier draft.

1 In fact in many civil cases one party admits fault and the dispute revolves around the amount of damages to be awarded. Sometimes the disagreement about quantum hinges on evidence: the court must determine facts (for example, how grave an injury is) in order to resolve it. If so, the civil standard of proof will still be needed in order to resolve this dispute. At other times the disagreement is legal rather than factual: the facts are agreed, but the disagreement is about the quantum merited by the factual situation; or about whether a factual situation can properly be categorised as, say, contributory negligence. In such situations analysis in terms of a standard of proof is less useful, because what is required is an evaluative judgment.

© The Modern Law Review Limited 1999 (MLR 62:2, March). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

happens, though, that there are a number of different conceptions of probability and that there is disagreement about precisely what probability means in the legal context.[2] This is not, however, a serious stumbling block for present purposes: whether forensic probability is cardinal or ordinal, objective or subjective, Pascalian or Baconian there is little doubt that it comes in degrees. This is certainly the accepted understanding of the difference between the criminal and civil standard of proof: in order to win a criminal case, the prosecution must prove its case to a higher degree of probability than need the plaintiff in a civil case. Though nothing much will turn on this, it will be simplest to interpret the concept of probability in law as requiring the fact-finder to have a certain degree of belief in a case before she accepts that it has been proved. Degrees of belief vary along a scale between 0 and 1; a person who has a degree of belief of 0.1 in a proposition believes that that proposition is unlikely to be true, whereas if her degree of belief is 0.99 she regards it as almost certainly being true.

The civil standard of proof has been described in a number of different ways. One of the clearest explanations has been offered by Lord Denning: '[i]f the evidence is such that the tribunal can say "we think it more probable than not", then the burden is discharged, but if the probabilities are equal it is not'.[3] Other formulations use expressions such as 'preponderance of probability', 'preponderance of the evidence' and 'balance of probabilities'. The second of these is the favoured term in the United States, while the third is the usual English formulation. All seem to imply the same thing: that the plaintiff's case need not be very probable, but that it needs to be more probable than its negation. This suggests that the best interpretation is that a degree of belief of just over 0.5 is required.[4] The only phrase which might be interpreted differently is the expression 'preponderance of the evidence', which might be thought to imply that the plaintiff needs to produce more evidence than the defendant, but this interpretation does not appear to have any current support. Indeed, it would be odd if it did, because it would imply that there was some way of quantifying evidence, as opposed to the effect that evidence has on belief.

2 See Colin Howson, 'Theories of Probability' (1995) 46 *British Journal for the Philosophy of Science* 1; Philip Dawid, 'Appendix: Probability and Proof' in Terence Anderson and William Twining, *Analysis of Evidence* (London: Weidenfeld and Nicolson, 1991) 404–416; William Twining, 'Debating Probabilities' (1980) 2 Liverpool L Rev 51.

3 *Miller* v *Minister of Pensions* [1947] 3 All ER 372, 373–374. Although Lord Denning states that it is 'well settled' that this is a correct description of the civil standard of proof, this passage has become the standard citation for the assertion that the civil standard of proof is proof on the balance of probabilities; before this, the law does not seem to have been clear and it is difficult to say when the balance of probabilities became the dominant concept. Although we find the concept of founding a verdict 'upon that which is most probable' as early as 1571 (*Newis* v *Lark* Plowd 403, 412), we should be wary of reading too much into statements like this one, made at a time when 'probability' may have had a rather different meaning than it does today. *Newis* certainly did not settle the civil standard: in 1858 we find the House of Lords, while declaring that 'in civil cases the preponderance of probability may constitute sufficient ground for a verdict', also noting that 'a difference of opinion exists on this question' (*Cooper* v *Slade* (1858) 6 HL Cas 746, 772). My suspicion is that, because the civil standard of proof may require a more sophisticated notion of probability than does the criminal standard, the modern concept only became established at a fairly late date. Although Barbara J. Shapiro's *Beyond Reasonable Doubt and Probable Cause: Historical Perspectives on the Anglo-American Law of Evidence* (Berkeley: University of California Press, 1991) concentrates on the development of the criminal standard, some of the sources she cites suggest that the courts did not differentiate between proof in civil and criminal cases until well into the 19th century (see eg, *ibid*, 27, 38). Indeed, in this century we find judges professing not to understand how there can be more than one standard of proof (eg Lord Goddard in *R* v *Hepworth and Fearnley* [1955] 2 QB 600, 603).

4 It might be argued that 'preponderance of probability' leaves open the possibility that the plaintiff's case must be substantially more probable than its negation. But if this is so, the phrase is too vague to serve as a description of a standard of proof.

© The Modern Law Review Limited 1999

The descriptions of the civil standard of proof, then, suggest that, for the plaintiff to win, the fact-finder needs to have a degree of belief of just over 0.5 in her case. But why 0.5 rather than some other degree of belief? What is so special about the balance of probabilities rule? It turns out that there are a number of reasons for adopting the 0.5 rule in civil cases, though they have not been discussed by the English courts.

One advantage of the 0.5 rule is that it minimises the expected number of errors made by the fact-finder.[5] This is not the same as minimising the *actual* number of errors,[6] but we are unlikely ever to know the actual error rate in the trial process, so minimising expected errors is the best we can attain. Nor is expected error minimisation the same as spreading errors equally over plaintiffs and defendants, but this is neither a sensible[7] nor – again, owing to our limited knowledge of actual error rates – attainable goal. However, if there is some positive correlation between the degrees of belief that fact-finders have in the cases presented to them and the actual truth or falsity of those cases (and, given that we are generally successful at navigating our way around the world, which involves constantly drawing inferences from evidence, this appears a plausible assumption), then the 0.5 rule will tend to minimise actual errors over the long run.

But what is so good about expected error minimisation? After all, that is not our goal in criminal cases; there, the higher standard of proof leads us to expect that we are making more errors than we might otherwise do. It is well known that we adopt a higher standard of proof in criminal cases because we are concerned to avoid errors of a certain kind: those that would lead to innocent defendants being found guilty. This suggests that it is our awareness of the expected consequences of our fact-finding decisions that drives our choice of the appropriate standard of proof.

This insight can be presented formally, using some basic decision theory.[8] To make a decision under risk[9] we need to consider the expected utility of that decision. The expected utility of a verdict for the plaintiff in a civil case, $EU(\Pi^v)$, is:[10]

$$P(\Pi^d) \times U(\Pi^v|\Pi^d) + P(\Delta^d) \times U^-(\Pi^v|\Delta^d)$$

$EU(\Delta^v)$, the expected utility of a verdict for the defendant, is:

$$P(\Delta^d) \times U(\Delta^v|\Delta^d) + P(\Pi^d) \times U^-(\Delta^v|\Pi^d)$$

5 See David Kaye, 'Naked Statistical Evidence' (1980) 89 Yale LJ 601, 604–605.

6 For a detailed discussion of the relationship between error rates and standards of proof, see Michael L. Dekay, 'The Difference Between Blackstone-Like Error Ratios and Probabilistic Standards of Proof' (1996) 21 Law & Soc Inq 95. As Dekay points out, the error rate depends not only on the standard of proof, but also on fact-finding accuracy and the proportion of deserving plaintiffs among cases tried.

7 As Kaye, n 5 above, 607, notes: '[m]istakes do not cancel one another out: it is no solace to the defendant who should have prevailed but did not that somewhere there is or will be a similarly affected plaintiff'.

8 See generally D.V. Lindley, *Making Decisions* 2nd ed (London: John Wiley, 1985). For applications of decision theory to adjudicative fact-finding, see J. Kaplan, 'Decision-Theory and the Factfinding Process' (1968) 20 Stanford L Rev 1065; Richard O. Lempert, 'Modeling Relevance' (1977) 75 Michigan L Rev 1021; David Kaye, 'The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation' (1982) 2 Am B Found Res J 487; D H. Kaye, 'Clarifying the Burden of Persuasion: What Bayesian Decision Rules do and do not do' (1999) 3 Int J Ev & Pr 1.

9 On the distinction between decision-making under risk and other types of decision-making, see Frederic Schick, *Making Choices: A Recasting of Decision Theory* (Cambridge: Cambridge University Press, 1997) 32–48.

10 This reads as follows: 'the probability that the plaintiff deserves to win × the utility of a verdict for the plaintiff given that the plaintiff deserves to win + the probability that the defendant deserves to win × the disutility of a verdict for the plaintiff given that the defendant deserves to win'. Concepts

© The Modern Law Review Limited 1999

If we assume that the utility of a plaintiff's verdict when the plaintiff deserves to win is the same as the utility of a defendant's verdict when the defendant deserves to win, and that the disutility of a plaintiff's verdict when the defendant deserves to win is the same as the disutility of a defendant's verdict when the plaintiff deserves to win, we find that the 0.5 rule maximises expected utility (because, under this assumption, $EU(\Pi^v) \geq EU(\Delta^v)$ if and only if $P(\Pi^d) \geq 0.5$).[11] An understanding of the expected utility of the possible decisions in a civil case, then, gives a further justification for the 0.5 rule, so long as we make the assumption, as above, that $U(\Pi^v|\Pi^d) = U(\Delta^v|\Delta^d)$ and that $U^-(\Pi^v|\Delta^d) = U^-(\Delta^v|\Pi^d)$. Expected utility is the second basic concept we need in order to understand standards of proof.[12]

At this point, it is worth pausing to gain further insights from the utility calculations presented above. Note that to calculate the expected utility of a decision for either the plaintiff or the defendant, we need to consider two utilities. For example, for a plaintiff's decision we need to know both $U(\Pi^v|\Pi^d)$ and $U^-(\Pi^v|\Delta^d)$. And to arrive at the appropriate standard of proof we need to know the $P(\Pi^d)$ for which $EU(\Pi^v) = EU(\Delta^v)$, which means that we need to assess four utilities.[13]

A second point to note is that, if we relax the assumption of symmetry made in the preceding paragraph, the 0.5 decision rule no longer maximises expected utility. Consider the utilities in a criminal case. The formulae above can still be used, all we need to do is read 'prosecution' instead of 'plaintiff' for 'Π'. Obviously, it is no longer appropriate to assume that $U^-(\Pi^v|\Delta^d) = U^-(\Delta^v|\Pi^d)$; the acceptance that these utilities are not symmetrical is one of the main features that distinguishes criminal from civil litigation. If we assume that $U^-(\Pi^v|\Delta^d) = 10 \times U^-(\Delta^v|\Pi^d)$ then we find that another, higher, standard of proof maximises expected utility.[14]

The presentation of standards of proof in terms of a calculation involving expected utility might appear to be inappropriate.[15] In criminal cases, an innocent defendant has what Dworkin refers to as a moral right not to be convicted;[16] our utility calculations might appear coolly to balance this right against broad policy concerns. In civil cases also, rights are involved. A deserving plaintiff has a legal right to damages, while a defendant who has broken no rules has the right to keep her money. But while the consequentialist terminology that I have adopted might give rise to some suspicions, a close analysis of just what is involved in the

---

such as '$(\Pi^v|\Pi^d)$' may require some further explanation. I am assuming that the world is in one of two possible states: that something happened such that the plaintiff deserves to win, or that something happened such that the defendant deserves to win. For example, either the defendant was negligent or she was not. If she was, and if negligence is the only contested issue, then the plaintiff deserves to win. If she was not, then the defendant deserves to win.

11 A general equation for deriving a standard of proof from the relevant utilities is given in Dekay, n 6 above, 111.

12 I am not suggesting that these are the only concepts that are necessary for a full understanding of forensic proof. In particular, I believe that a concept of weight or resilience is also necessary. This concept will not, however, play a significant role in the discussion in this article.

13 Certain authors who have used decision theory to analyse standards of proof have assumed that standards of proof can be set by considering only two utilities: the disutilities of mistaken verdicts. However, because the assumptions that allow this simplification are difficult to justify (either that $U(\Pi^v|\Pi^d)$ and $U(\Delta^v|\Delta^d) = 0$; or that $U(\Pi^v|\Pi^d) + U^-(\Pi^v|\Delta^d) = 0$ and that $U(\Delta^v|\Delta^d) + U^-(\Delta^v|\Pi^d) = 0$) I shall not make them here. See Dekay, n 6 above, 115–117.

14 If either of the simplifying assumptions described in the preceding footnote are made, the standard of proof would be just over 0.9.

15 For criticism, see Andrew Ligertwood, 'The Uncertainty of Proof' (1976) 10 Melb L Rev 367, 368–375.

16 Ronald Dworkin, 'Principle, Policy, Procedure' in *A Matter of Principle* (Oxford: Clarendon Press, 1985).

© The Modern Law Review Limited 1999

calculations reveals that there is no vulgar utilitarian balancing going on. Dworkin's argument is that procedural rules – which include rules setting standards of proof – should respect two principles: first, that correct importance is attached to the risk of moral harm; secondly, that the risk of moral harm is consistently weighted over a set of similar cases.[17] There must, however, be some balancing of the risk of moral harm against other values (trial costs, effective enforcement of the law, and the like) before these two principles come into play. This is why the defendant in a criminal case does not have the right to the most accurate procedures possible (for example, the highest conceivable standard of proof), but only to procedures which obey the two principles above, even though this means that innocent defendants risk being convicted, and thus having their moral rights infringed. The same applies in civil cases, but here the initial balancing involves balancing, among other things, the plaintiff's right to a deserved verdict against the defendant's right to a deserved verdict.[18] Our utility formulae are best understood as capturing just this balancing process. However, the doubts raised about them in this paragraph do prompt an important question: can the right to consistent weighing of moral harm be protected if the determination of the relevant utilities is left to each individual fact-finder? If it is, the principle of consistency may be infringed, because there is no guarantee that different fact-finders in similar cases will adopt the same utilities. This will become an important issue later, where it will be dealt with more fully.

So far we have seen that the 0.5 rule minimises expected errors and that, as a direct result of this, where $U(\Pi^v|\Pi^d) = U(\Delta^v|\Delta^d)$ and $U^-(\Pi^v|\Delta^d) = U^-(\Delta^v|\Pi^d)$ it will also maximise expected utility. I have defended this means of conceptualising standards of proof against accusations of vulgar utilitarianism, but now another question looms: why assume symmetrical utilities?

There are good reasons for making this assumption in the typical civil case. In the typical case, where the plaintiff is alleging that the defendant has broken some rule, she is claiming a legal right to compensation. If she is the victim of a mistaken verdict, then her rights are infringed; but if she benefits from a mistaken verdict then the defendant's rights are infringed. There will usually be no reason for valuing the defendant's rights more than the plaintiff's rights; consequently, there is no reason for preferring an error in one direction to one in the other. Fact-finding in the typical civil case should therefore be governed by a principle of equality, and this provides a justification for the 0.5 rule. There are two dimensions to this principle – which I shall term the 'equality principle' – which reflect Dworkin's two principles for procedural rules. If, in some case, a court were to raise the standard of proof by which a plaintiff had to prove her case, it would be implicitly stating that it regarded an error adversely affecting the plaintiff as preferable to an error adversely affecting the defendant (a violation of the first dimension). But the court would also suggest that it regarded errors adversely affecting the plaintiff as less serious than similar errors in other cases, where the usual 0.5 standard was applied (a violation of the second dimension). Thus the 0.5 rule upholds equality

17 *Ibid*, 89. The way in which I have expressed this second principle involves some slightly speculative reading of Dworkin's essay. At 92–93, where his discussion turns to civil procedure, he expresses the second right as a 'right to a consistent evaluation of [moral] harm in the procedures afforded them as compared with the procedures afforded others in different civil cases'. I presume he here means different cases of the same type, rather than all civil cases.

18 The balancing process involves more than just balancing these two rights against each other, because the more elaborate the fact-finding procedures adopted, the more likely the deserving party is to receive her deserved verdict. Thus the utility of correct verdicts must be balanced against the cost of elaborate procedures.

© The Modern Law Review Limited 1999

between plaintiffs and defendants, while consistent application of a standard of proof (though not necessarily the 0.5 standard) upholds equality between plaintiffs as a group.[19]

There might, of course, be reasons for departing from this principle of equality where there are grounds for regarding the various utilities as asymmetrical. But before considering what sorts of cases these might be, we need to survey some possible objections to the justification of the 0.5 standard in terms of equality. The first objection is founded on the observation that, in practice, the legal system does not treat plaintiffs and defendants equally. The most obvious departure from the equality principle is the way in which ties are treated in civil cases: if the plaintiff's case and its negation are equiprobable, the plaintiff loses. To win, her case must be more probable than not ($P(\Pi^{\alpha})$ must be >0.5). I do not, however, consider this to be a very serious objection. In theory, a successful plaintiff may have established her case by as small an amount as you like above 0.5. A case established to 0.5 0000001 is still a plaintiff's case, and here the departure from equality is minimal. Some rule is needed to deal with ties and, although a rule awarding equiprobable cases to plaintiffs would function just as well (as would tossing a coin to decide ties), the >0.5 rule fits better with the general rule placing the burden of proof on the plaintiff.[20]

It is just this practice of placing the burden of proof on plaintiffs, though, which seems to point to a larger violation of the equality principle. In many cases, the burden of proof (the 'evidential burden' or 'burden of production' as it also called) will have more effect on who wins or loses than will the standard of proof applied. Even if standards of proof were designed to favour plaintiffs and were set at, say, 0.25, a plaintiff who could produce no evidence and who bore the burden of proof would still lose.[21] Admittedly, the burden of proof rules require explanation, but their explanation need not substantially undermine the equality principle. The rules

19 I should point out that certain commentators do not agree that expected error minimisation and equality between litigants provide a sound justification for the 0.5 rule. Those writing from a law and economics perspective tend to argue that the civil standard of proof should be set at a level which will ensure optimal deterrence of tortious conduct (ie it should not under-deter, increasing the number of accidents, but nor should it over-deter, increasing the cost of safety measures and encouraging potential victims to be careless). See Richard A. Posner, *Economic Analysis of Law* 3rd ed (Boston, Little Brown, 1986) 550–552; Jason S Johnston, 'Bayesian Fact-finding and Efficiency. Toward an Economic Theory of Liability Under Uncertainty' (1987) 61 S Cal L Rev 137. For my present purposes, I find little of value in this literature, first, because it does not fit into the rights-based analysis which informs my approach to the problem; secondly, because even if one could determine what level of deterrence is optimal, determining the standard of proof which would achieve this level (which would require knowledge about the accuracy of verdicts and about the proportion of deserving cases among those that come to trial) appears to be an extraordinarily difficult task.

A rather different proposal is made by McCarthy, who suggests that we should 'vary the standard of proof with the kind of evidence that can be expected' (David McCarthy, 'Liability and Risk' (1996) 25 Phil & Pub Aff 238, 258). While this is an interesting suggestion, his analysis is marred by ambiguities in and misinterpretations of the probabilities in the example he uses; for example, at 259 he presumes that a corollary of the 0.5 rule is that 'we tolerate a large number of cases in which the wrong person will pay compensation'.

20 In *Rhesa Shipping Co SA* v *Edmunds* [1985] 2 All ER 712, 718, Lord Brandon remarked that '[n]o judge likes to decide cases on burden of proof if he can legitimately avoid having to do so. There are cases, however, in which, owing to the unsatisfactory state of the evidence or otherwise, deciding on the burden of proof is the only just course for him to take'. It may be that this reluctant attitude reflects an intuition that it is rather unfair to decide against a plaintiff whose case has been established to the same degree of probability as the defendant's, because this violates the equality principle.

21 The amount of evidence a plaintiff needs to produce to discharge her evidential burden is determined by the standard of proof: the lower the standard, the easier it is to discharge the burden. See James McNaughton, 'Burden of Production of Evidence: A Function of a Burden of Persuasion' (1955) 68 Harv L Rev 1382; *Anderson* v *Liberty Lobby, Inc* 477 US 242 (1986) 252–256.

© The Modern Law Review Limited 1999

might be better thought of as restricting the class of people who are entitled to call themselves plaintiffs, rather than as operating to make the task of plaintiffs more onerous. Although, in one sense, all that is required for a potential litigant to become a plaintiff is to issue a writ, if this were really all that was required to achieve plaintiff status the courts might be overwhelmed with cases, and many faultless defendants might settle cases out of court rather than go to trial.[22] By placing the burden of proof on plaintiffs, potential litigants whose cases lack merit are deterred from issuing writs. Another explanation of the burden of proof rules that does not involve regarding them as violations of equality – and which is not unlike the previous explanation – is that they reflect basic principles of civility.[23] We do not generally accuse people of doing us wrong unless we are able to back up the explanation with evidence, and in law this requirement is fairly easily mapped onto the rules governing the burden of proof.[24]

A third objection to the equality principle builds on observations of the workings of the legal system, such as those discussed above, but goes somewhat further, to claim that courts prefer returning a wrong verdict against a plaintiff, and denying her damages she does deserve, to returning a wrong verdict against a defendant, and thus giving the plaintiff damages she does not deserve (in terms of utilities, $U^{-}(\Delta^{v}|\Pi^{d}) > U^{-}(\Pi^{v}|\Delta^{d})$.[25] On this analysis, the legal system is, and perhaps should be, biased towards the status quo.[26] If this were true, it would not only explain the >0.5 rule and the practice of placing the burden of proof on the plaintiff, but would also justify applying a standard of proof of substantially more than 0.5 – 0.6, perhaps – to the plaintiff's claim, to guard against decisions where

22 Defendants will tend to settle for any amount smaller than their expected loss at trial. Even a defendant who judges her chances of winning to be high would be prepared to settle for a small amount to avoid both the certain loss of non-recoverable litigation costs (stress, time etc) and the small chance of losing and having to pay damages and indemnify the other party.

23 See Dale A. Nance, 'Civility and the Burden of Proof' (1994) 17 *Harvard Journal of Law and Public Policy* 647.

24 There are a number of other analyses of how proof burdens should be allocated. See, for example, Bruce L. Hay, 'Allocating the Burden of Proof' (1997) 72 Indiana LJ 651; Bruce L. Hay and Kathryn E. Spier, 'Burdens of Proof in Civil Litigation: An Economic Perspective' (1997) XXVI J Leg Stud 413; Thomas R. Lee, 'Pleading and Proof: The Economics of Legal Burdens' [1997] Brigham Young U L Rev 1. These analyses provide a general justification for placing the initial burden of proof on the plaintiff.

25 By which I mean that its *negative* value is greater.

26 A claim made by certain commentators, though rarely defended at any length. See, for example, Alex Stein, 'The Refoundation of Evidence Law' (1996) IX Can J Law and Juris 279, 335, where the >0.5 rule is justified on the grounds that '"taking" is perceivable as being generally more harmful than "not giving"'. Stein does not, however, build this bias into his general analysis of the civil burden, claiming, at 334, that the 0.5 rule 'minimize[s] the overall damage to be incurred, unavoidably, by the legal system', which suggests that he does not take this point seriously. See also David T. Wasserman, 'The Morality of Statistical Proof and the Risk of Mistaken Liability' (1991) 13 Cardozo L Rev 935, 937 ('[e]ven in civil cases, higher odds may be required to upset than preserve the status quo because of the greater disutility of judicially imposed error'); Michael D. Bayles, *Procedural Justice: Allocating to Individuals* (Dordrecht: Kluwer, 1990) 115–139. Bayles is not careful to separate ordinary civil litigation from administrative fact-finding, so it is not clear whether he would apply his belief that standards of proof should be biased against allocating to the undeserving to civil litigation. For the view that the >0.5 rule is not optimal, see Gordon Tullock, *Trials on Trial: The Pure Theory of Legal Procedure* (New York: Columbia University Press, 1980) 83–84, 86. Unfortunately, Tullock's reasoning (which involves the assumption that the >0.5 rule will result in the rejection of more just claims than the 0.5 rule) is flawed.

Some commentators make the point that the enforcement costs incurred when there is a judgment for the plaintiff might justify disfavouring plaintiff's verdicts; see Posner, n 19 above, 552; Lee, n 24 above. While this observation might be used as a reason for raising the standard of proof, it would not be a compelling one. As Lee notes (at 12), enforcement costs are extremely variable: sometimes they are huge, sometimes negligible. Incorporating them into the calculation of proof standards would therefore be difficult; it would also lead to differential treatment of plaintiffs.

© The Modern Law Review Limited 1999

the court will mistakenly disturb the status quo. However, the objection that disturbing the status quo carries extra disutility ignores the fact that if the plaintiff deserves to win, the status quo has already been disturbed.[27] Taking the term 'status quo' to refer to the point in time before the alleged wrong occurred, the failure to right the wrong is just as bad as the mistaken imposition of a burden on the defendant. Although it may sometimes be correct to say that the failure to right a wrong is not as bad as doing that same wrong, it is not obviously correct to say this in the context of a judge's 'role omission' – the failure to act when there is a duty to act.[28] Further, from the litigant's point of view it appears to make little difference whether she loses £300 through Z's negligence, or through the mistaken decision of a judge. If we have a lingering intuition that the latter is worse it is due, I suspect, to our seeing the person who is the victim of a mistaken judicial decision as some complete innocent, plucked off the street at random and made to pay damages to a stranger. This will rarely be the case. More usually that person will have chosen to engage in some activity – say driving, or entering into a contract – that carries with it the risk of litigation, and hence of the mistaken imposition of liability.[29]

The justifications for the 0.5 rule outlined above establish the importance of the rule, at least so long as the utilities used in the fact-finder's decision-making are taken to be symmetrical. In a later part of this article we will consider situations where asymmetries in these utilities may justify some variation in the standard of proof, so it is worth, at this point, underlining that there are good grounds for sticking to the 0.5 rule in the absence of strong reasons for departing from it. The rule not only reflects the two dimensions of the equality principle; it also minimises expected errors over the long run. The avoidance of expected errors is a good in itself, though one that may not easily be incorporated into the decision calculus in any individual case.[30] There should therefore be some preference for the 0.5 rule, even when asymmetrical utilities point towards raising the standard of proof so as to avoid an error against one party. Together with the two dimensions of the equality principle, this suggests that the 0.5 rule should be protected by a degree of inertia and that it should only be abandoned where the reasons for doing so are compelling.

Judicial interpretations of the civil standard of proof

The development of variations in the basic civil standard of proof has been driven by the fact that certain civil cases involve serious issues. Some involve serious allegations (of criminal conduct, for example); some may have serious outcomes (a

27 cf Sunstein's critique of 'status quo neutrality' in constitutional adjudication. Cass R. Sunstein, *The Partial Constitution* (Cambridge, Mass: Harvard University Press, 1993) esp 68–92, 127–133.

28 The phrase is Honderich's: see Ted Honderich, *Violence For Equality: Inquiries in Political Philosophy* revised ed (London: Routledge, 1989) 72. The validity of the act/omission distinction is, of course, much contested. But none of the factors commonly taken to distinguish omissions from acts appear to apply to the judge in our example. See Honderich, 64–103; Alison McIntyre, 'Guilty Bystanders? On the Legitimacy of Duty to Rescue Statutes' (1994) 23 Phil & Pub Aff 157.

29 It seems that where the defendant really does have a claim to have been plucked at random, people have a genuine, and perhaps justifiable, reluctance to find liability when the case is proved against her by just a whisker over 0.5. This may explain the common reaction to the 'gatecrasher' and 'blue bus' thought experiments (see Gary L. Wells, 'Naked Statistical Evidence of Liability: Is Subjective Probability Enough?' (1992) 62 *Journal of Personality and Social Psychology* 739). For an analysis which largely supports this view, see Wasserman, n 26 above; cf Richard D. Friedman, 'Answering the Bayesiokeptical Challenge' (1997) 1 Int J Ev & Pr 276, 279.

30 Utilitarianism is sometimes criticised because values such as truth often do not get weighted in utility calculations. See, for example, Peter Jones, 'The Ideal of the Neutral State' in Robert E Goodin and Andrew Reeve (eds), *Liberal Neutrality* (London: Routledge, 1989) 15.

© The Modern Law Review Limited 1999

parent losing custody of her child); and some involve both.[31] In such cases, the courts are reluctant to find a claim proved on the simple balance of probabilities, without at least noting that the gravity of the case requires some deviation from the normal approach to proof on this standard. When it comes to describing the modification of approach that is required, however, we find that judges adopt different approaches and that their explanations are sometimes unclear. In this section I examine some of the variations that have been proposed and the circumstances in which they have been applied.[32]

In *Miller* v *Minister of Pensions*, Lord Denning gave what is probably the clearest and most frequently cited description of the civil standard of proof in English law.[33] In *Bater* v *Bater*,[34] however, Lord Denning delivered a judgment which has probably done more than any other to produce confusion about the civil standard. In *Bater* the appellant, who had petitioned for divorce on the grounds of cruelty, was appealing against the decision of a commissioner who had dismissed her petition because cruelty had not been proved beyond reasonable doubt.[35] Dismissing the appeal, Lord Denning claimed that neither the civil nor the criminal standard of proof was an absolute standard. A civil case 'may be proved by a preponderance of probability', he explained, 'but there may be degrees of probability within that standard. The degree depends on the subject-matter. A civil court, when considering a charge of fraud, will naturally require a higher degree of probability than that which it would require if considering whether negligence was established'. In other words, the degree of probability required should be 'commensurate with the occasion'. This was not an altogether novel concept; as Denning noted, '[m]any great judges have said that, in proportion as the crime is enormous, so ought the proof to be clear'.[36] In civil litigation a similar view had been expressed by the Court of Appeal on previous occasions,[37] as well as by the High Court of Australia.[38] What has perhaps led to the longevity of Denning's comments in *Bater* is that adopting them finesses two of the more troubling aspects of a flexible standard of proof. Denning saw no inconsistency between the concept of a flexible standard of proof and its description in conventional terms, as proof by 'a preponderance of probability'. Further, while the civil standard of proof was held to be flexible, it was held not to be so flexible as to blur the important distinction between the civil and criminal standards of proof.[39] The upshot of this

31 The distinction between allegations and outcomes is made in Rosemary Pattenden, 'The Risk of Non-Persuasion in Civil Trials: The Case Against a Floating Standard of Proof' (1989) 8 CJQ 220.

32 For more complete reviews of the case-law, see *ibid*; Colin Tapper, *Cross and Tapper on Evidence* 8th ed (London: Butterworths, 1995) 158–172.

33 See n 3 above and accompanying text.

34 [1950] 2 All ER 458.

35 An indication of the gravity with which divorce was then regarded is provided by the comments of Bucknill LJ, *ibid* 458–459: 'I regard proceedings for divorce as proceedings of very great importance, not only to the parties, but also to the State. If a wife is divorced, she not only has the stigma on her for the rest of her life, but it may .. mean ruin to her'. This despite the fact that in this case it was the *wife* who was petitioning for divorce.

36 *ibid* 459.

37 *Gower* v *Gower* [1950] 1 All ER 804; *Davis* v *Davis* [1950] 1 All ER 40.

38 *Briginshaw* v *Briginshaw* (1938) 60 CLR 336.

39 A civil court, Denning stated, 'does not adopt so high a degree [of proof] as a criminal court, even when it is considering a charge of a criminal nature' (n 34 above, 459). Contrast the judgment of Bucknill J, who explicitly noted that, in an earlier case (*Gower* v *Gower* [1950] 1 All ER 804) 'I should, perhaps, have said: "The standard of proof required in a criminal case is higher than required in *some* civil actions"' (at 458, emphasis in original). Even Denning himself was perhaps not too sure of this aspect of his judgment; some years later, in *Hornal* v *Neuberger Products Ltd* [1956] 3 All ER 970, 973, while applying the notion of a flexible standard, he was to say that the degree of probability required in a civil case '*need* not ... reach the very high standard required by the criminal law' (my emphasis).

© The Modern Law Review Limited 1999

interpretation of the civil standard of proof was that the commissioner's reference to proof beyond reasonable doubt could not be criticised, because the degree of doubt which he would have thought reasonable would have been commensurate with the seriousness of the case.

Six years later, Lord Denning was sitting in the Court of Appeal when the nature of the civil standard of proof was again in issue. In *Hornal* v *Neuberger Products Ltd*[40] the question of fact was whether or not a representation had been made to the plaintiff that a lathe he had bought had been 'Soag reconditioned'. The issue had become complicated because, fraud being a possibility, the trial judge had been unsure as to the standard of proof – proof beyond reasonable doubt or on the balance of probabilities – which should be applied. All three appeal judges agreed that the appropriate standard of proof was the civil standard, and that the ease with which this standard could be satisfied varied from case to case. While Lord Denning simply observed that '[t]he more serious the allegation the higher the degree of probability that is required',[41] the judgments of Lords Hobson and Morris suggest a different approach. According to Lord Hobson, 'the balance of probability may be more readily tilted in one case than in another'.[42] Lord Morris' dicta on the issue are more ambiguous, but the one that has become best known is the comment that 'the very elements of gravity become a part of the whole range of circumstances which have to be weighed in the scale when deciding as to the balance of probabilities'.[43] Lord Hobson may be interpreted as suggesting that, rather than the standard of proof varying (as it does in Denning's scheme), it is the amount of evidence which is needed to satisfy the standard which differs from case to case. It is more difficult to interpret Lord Morris' comments on the civil standard of proof, but it may be that he was influenced by a theory of presumption under which a presumption is an item of evidence to be considered by the fact-finder.[44] There is certainly some support for this approach in earlier cases concerning the proof of crime in civil proceedings, where it had been held that, although the appropriate standard of proof was the civil standard, the presumption of innocence was still relevant.[45] On this interpretation, Lord Morris' analysis is closer to Hobson's than to Denning's, because it suggests that more evidence is required to prove more serious allegations. Indeed, this is how it was interpreted in a later case, *Re Dellow's Will Trusts*, where, after quoting Lord Morris, Ungoed-Thomas J stated that '[t]he more serious the allegation, the more cogent is the evidence required to overcome the unlikelihood of what is alleged and thus to prove it'.[46]

Bater and *Hornal* established the two main deviations from the basic civil standard of proof. I shall term the two approaches the 'flexible standard' (*Bater*) and the 'prior probability approach' (*Hornal*, as per Lords Hobson and Morris). The basic distinction between them is that under the former the standard of proof varies according to the seriousness of the issues, while under the latter the standard remains fixed but the degree of evidence needed to satisfy it varies because more

40 [1956] 3 All ER 970.
41 *Ibid*, 973.
42 *Ibid* 977.
43 *Ibid* 978.
44 See the comments of Lord Denning in *Stupple* v *Royal Insurance Co* [1971] 1 QB 50, 72.
45 See, for example, the comments of Patteson J in *Doe d Devine* v *Wilson* (1855) 10 Moo PC 501, 531. '[t]he jury must weigh the conflicting evidence, consider all the probabilities of the case not excluding the ordinary presumption of innocence and must determine the question according to the balance of those probabilities'.
46 [1964] 1 All ER 771, 773.

© The Modern Law Review Limited 1999

serious events are said to be less probable.[47] In terms of conceptual development, later cases have added little. Both approaches have been approved by the House of Lords: the flexible approach in *Blyth* v *Blyth*,[48] and the prior probability approach in *Re H and others (minors)*.[49] The latter decision is significant. A series of cases on child abuse had established that in such cases some variation in the standard of proof was necessary, with both approaches finding some favour.[50] In *Re H*, a majority of the House of Lords decided that the prior probability approach should be taken. Or, as it was put by Lord Nicholls, the standard of proof was the balance of probabilities, but '[w]hen assessing the probabilities the court will have in mind as a factor ... that the more serious the allegation the less likely it is that the event occurred and, hence, the stronger should be the evidence before the court concludes that the allegation is established on the balance of probability'.[51] The decision of the House of Lords is probably only authority for the standard of proof in child abuse cases (and possibly only for cases involving proof under section 31(2) of the Children Act 1989).[52] It seems that the flexible standard is alive and well, judging by the Court of Appeal's comments in *R* v *Secretary of State for the Home Department, ex parte Rahman*.[53]

This sketch of judicial approaches to the civil standard of proof is complicated by the existence of a third approach. In *R* v *Hepworth and Fearnley*[54] Lord Goddard CJ declared that he was unable to understand the difference between the criminal and civil standards of proof. This may imply that in both cases the fact-finder must be satisfied, or believe, or perhaps be certain that the claim of the party bearing the burden of proof is true before returning a verdict in her favour.[55] At one time this view had considerable support in Australia,[56] but it has rarely been

47 In a few cases, one can find a variation on the prior probability approach. It is sometimes suggested that what is required in serious cases is not more evidence, but more careful scrutiny of the evidence. See, for example, *Brady (Inspector of Taxes)* v *Group Lotus Car Companies plc* [1987] 3 All ER 1050, 1060; *N. Michalos & Sons Maritime SA* v *Prudential Assurance Co Ltd* [1984] 2 Lloyd's Rep 264, 271. While initially appealing, this approach might imply that judges need not be too careful when scrutinising evidence in less serious cases.

48 [1966] AC 643.

49 [1996] 1 All ER 1.

50 *Re M (A Minor)* [1994] 1 FLR 59 (prior probability approach); *R* v *Birmingham Juvenile Court, ex p S* [1984] Fam 93; *Re G (A Minor)* [1987] 1 WLR 1461; *H* v *H (Minor), K* v *K (Minors)* [1989] 3 WLR 933; *Re W (Minors)* [1994] 1 FLR 419; *Re L* [1996] 1 FLR 116 (flexible standard).

51 n 49 above, 16. Lords Goff and Mustill gave judgments concurring with Lord Nicholls. Lord Lloyd expressly disagreed with Lord Nicholls on this point, preferring to apply the basic standard of proof without adding any comments on the cogency of evidence required to satisfy it. Lord Browne-Wilkinson expressed agreement with Lord Nicholls that the standard was the ordinary civil standard, but did not add any comments about the amount of evidence required.

52 Lord Lloyd explicitly confined his comments to the section of the Act under consideration. In the Court of Appeal, however, Millett LJ asserted that '[i]n civil cases ... there is only one standard of proof: proof on the balance of probabilities. It is never necessary to prove facts to a standard beyond the balance of probabilities' ([1995] 1 FLR 643, 659). Considering the decision of the House of Lords in *Blyth* v *Blyth* this was a rather incautious comment. But cf Ward LJ's claim in *J* v *S-T (Formerly J) (Transsexual: Ancillary Relief)* [1998] 1 All ER 431, 452: '[i]n [*Re H*] the House of Lords was unanimous in its opinion that the standard of proof in all civil proceedings is the ordinary civil standard of proof on a balance of probabilities'.

53 [1997] 1 All ER 796.

54 [1955] 2 QB 600, 603.

55 Some dicta where this view is expressed, however, are open to the interpretation that what is being applied is a version of the flexible standard of proof, with 'satisfied' denoting the belief state where, given the seriousness of the case, one is prepared to say that it has been proved.

56 The clearest expression of this approach is to be found in the judgment of Sir Owen Dixon in *Briginshaw* v *Briginshaw* (1938) 60 CLR 336, 361. '[t]he truth is that, when the law requires the proof of any fact, the tribunal must feel an actual persuasion of its occurrence or existence before it can be found. It cannot be found as the result of a mere mechanical comparison of probabilities

© The Modern Law Review Limited 1999

endorsed by English judges. A possible exception is Lord Scarman, who in *R* v *Home Secretary, ex parte Khawaja*[57] referred to the need to be satisfied.[58] The distinguishing feature of the belief/satisfaction approach is its rejection of the notion of probability: if the judge is uncertain, she must decide the case against the party bearing the burden of proof. This view finds support in the work of Nesson, who has claimed that, ideally, verdicts should be statements about belief, not about probability.[59]

Having gained an overview of judicial approaches to the civil standard of proof, it is necessary to fill out this picture with some consideration of how the variations on the basic standard have been applied in practice and of the problems they have caused. A preliminary point is that the distinction between the basic standard and the varied standard (of whichever type) is usually taken seriously. Appeals are sometimes allowed because trial judges have applied the normal civil standard in circumstances which the appellate court holds to require a varied standard.[60] In civil cases heard by a jury, however, the variation is not always insisted on.[61] This is problematic, because it appears to involve a breach of the second dimension of the equality principle. Alternatively, the judge can attempt to explain the special standard of proof to the jury, but then the linguistic problems, which were simply ignored by Lord Denning in *Bater*, are significant. In *Slattery* v *Mance*, for example, the jury was told that: '[t]his is a serious case and it may be although this onus is only to satisfy you on the balance of probability that you, as a jury, consider as this is arson ... you would require a high degree of probability before you would say it has been made out that he fired the ship'.[62] This runs the risk of confusion: 'balance' suggests equiprobability whereas 'high degree' does not.

independently of any belief in its reality'. Other Australian cases where one finds a similar approach are discussed in Australian Law Reform Committee, *Evidence (Interim Report)* vol 2 (1985) 321–325; Sir Richard Eggleston, *Evidence, Proof and Probability* 2nd ed (London: Weidenfeld and Nicolson, 1983) 129–140; David Hamer, 'The Civil Standard of Proof Uncertainty: Probability, Belief and Justice' (1994) 16 Sydney L Rev 506, 509–512.

57 [1984] 1 AC 74. This was a decision of the House of Lords, and its status is now unclear. It was not referred to at all in *Re H* and thus has caused some confusion. see *Inland Revenue Commissioners* v *Botnar* [1998] STC 38, 74.

58 It is not easy to ascribe a clear view to Lord Scarman from his comments in *Khawaja* However, at 113 he remarks that it is difficult to distinguish the civil and criminal standard, asking '[i]f a court has to be satisfied, how can it at the same time entertain a reasonable doubt ...?'. At 114 he states that 'the degree of probability must be such that the court is satisfied'. These comments suggest that his views are similar to those of Lord Goddard.

59 Charles Nesson, 'The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts' (1985) 98 Harv L Rev 1357, 1362. Because it finds so little support in English law, I will not consider this approach in detail. For criticism of Nesson's view, see Ronald J Allen, 'Rationality, Mythology and the "Acceptability of Verdicts" Thesis' (1986) 66 Boston U L Rev 541. My own view is that the approach is untenable because it runs aground on the problem of deductive closure. Suppose a judge who takes the 'satisfaction' view has returned plaintiffs' verdicts in 300 cases. In each case, she is satisfied that the plaintiff's account was true. Now imagine asking this judge whether she is satisfied that she was not mistaken in any of those cases. It would take a very brave judge to reply 'yes'. Yet if her reply is 'no' she seems to be violating the principle of logical consistency under deductive closure, because she implies that in at least one case she was not satisfied. Readers familiar with the philosophical literature on belief will recognise this as a version of the paradox of the preface: see Mark Kaplan, *Decision Theory as Philosophy* (Cambridge: Cambridge University Press, 1996) 102–154; Daniel Hunter, 'On the Relationship Between Categorical and Probabilistic Belief' (1996) 30 *Noûs* 75.

60 The best example is *Re W (Minors)* [1994] 1 FLR 419, 426, where Balcombe LJ remarks: 'I have to say that I am unhappy about the judge's findings on this issue. The very fact that the judge did not feel confident about the precise forms of sexual abuse suffered by each girl suggests that if, as I believe, he should have applied the higher standard of proof, he might not have been satisfied'. See also *Serio* v *Serio* (1983) 4 FLR 756.

61 *Lawrence* v *Chester Chronicle, The Times* 8 February 1986.

62 [1962] Lloyd's Rep 60, 63.

© The Modern Law Review Limited 1999

A second point is that there is a fair degree of confusion about whether there is a real distinction between the two approaches. The implication of the House of Lords' decision in *Re H* is that there is a difference. This is surely right: one approach raises the standard of proof; the other is an observation about the quantum of evidence required to satisfy the standard of proof. Yet the distinction has been described as 'academic'[63] and as 'one of expression only'.[64] On occasion, such confusion may be significant. In *Re W (Minors)*, Balcombe J speaks of a higher standard of proof (which suggests the flexible approach), but justifies adopting it in terms of the prior probability of abuse.[65] The question in that case was whether a child had been abused, and the issue was raised whether an allegation that the father had abused the child required a higher standard of proof than the allegation that the child had been abused.[66] The precise approach which is taken here is important because, on the prior probability approach, once it is proved that the child has been abused the prior probability that the father was the culprit is presumably quite high; whereas if the allegation of patriarchal abuse ratchets up the standard of proof, that allegation will be more difficult to prove. Consider also *Re M (A Minor)*,[67] where the Court of Appeal dismissed a local authority's appeal against, among other things, the trial judge's finding that M had not suffered serious physical abuse. The trial judge had applied the flexible standard of proof, prompting the Court of Appeal to remark that the prior probability approach was preferable. But the trial judge may well have considered the prior probability of abuse to be low; if so, he probably took this into account before applying (as he did) the higher standard of proof. If the Court of Appeal thought that the standard of proof should have been the simple balance of probabilities, it should have paid less deference to the trial judge's findings of fact because he had, essentially, made an error of law. The judge's mistake was not compensated for by the Court's noting the prior improbability of abuse and, as this was a case where the evidence was fairly evenly balanced, the difference might have affected its outcome.

There is also some confusion about just what it is that must be serious before the higher standard of proof is deemed appropriate. The case law suggests that a case will be deemed serious when a finding of fact may stigmatise the defendant in some way,[68] but it is not obvious how courts determine whether a finding will involve stigma. For example, in *Hornal* the trial judge considered that if the plaintiff's claim was made on the basis of contractual warranty, the ordinary civil standard should apply, but that if the claim was put on fraud, a higher standard was required. If the seriousness of the allegation is what is held to be important, this seems reasonable: fraud has criminal connotations whereas breach of warranty does not. Lord Denning, however, held that the same (higher than normal) standard should be applied however the claim was put, because 'the defendant's moral guilt is just as great whatever the form of action'.[69] In *F v F (Divorce: Insolvency: Annulment of Bankruptcy Order)*,[70] the 'deceitful presentation' of the husband's financial circumstances was held to require a higher standard of proof. These

63 Per Ungoed Thomas J, in *Re Dellow's Will Trusts*, n 46 above, 773.
64 Per Waite LJ, in *Re M*, n 50 above, 67.
65 [1994] 1 FLR 419, 425.
66 This distinction had been made in *Re G (A Minor)* [1987] 1 WLR 1461.
67 [1994] 1 FLR 59.
68 Or another person connected with the proceedings; see *Preston Jones* v *Preston Jones* [1951] AC 391, where the House of Lords was concerned about 'bastardising' a child.
69 n 40 above, 973.
70 [1994] 1 FLR 359. See also *Baker* v *Baker* [1995] 2 FLR 829.

© The Modern Law Review Limited 1999

decisions suggest that any judicial finding of dishonesty, or perhaps even recklessness, will stigmatise a defendant, who should therefore be protected by a higher standard of proof. Yet if this were taken seriously, the higher standard of proof would be applied in any case where there is a conflict of evidence which cannot be explained on grounds of the defendant's mistake, because a judgment for the plaintiff will effectively brand the defendant a perjurer.

Another important point is that the higher standard of proof has been imposed in cases where the standard of proof is specified by legislation. In *Serio* v *Serio*,[71] The Court of Appeal considered section 26 of the Family Law Reform Act 1969, which reads: 'Any presumption of law as to the legitimacy or illegitimacy of any person may in any civil proceedings be rebutted by evidence which shows that it is more probable than not that that person is illegitimate'. This section abrogated the old rule that proof beyond reasonable doubt was required to rebut the presumption of legitimacy.[72] The Court held that the trial judge had erred by applying the simple civil standard of proof to the question of legitimacy; he should instead have applied 'a standard commensurate with the seriousness of the issues involved'.[73]

Finally, we should note that the problem of whether the basic civil standard of proof is appropriate in serious civil cases arises in other common law jurisdictions. In Australia, the flexible standard of proof has recently been enshrined in legislation.[74] Canada also recognises the flexible standard.[75] In the United States, however, the flexible standard is not recognised. Instead, a single intermediate standard of proof is used to deal with some serious issues in civil and administrative hearings.[76] This third standard of proof is the requirement of proof by 'clear and convincing evidence'.[77]

Evaluating judicial approaches to the civil standard of proof

Having sampled various judicial approaches to the civil standard of proof, we are now able to use the conceptual framework previously developed to evaluate them. The two most common variations on the ordinary civil standard are the flexible approach and the prior probability approach; I shall examine the cogency of each in some detail.

---

71 (1983) 4 FLR 756.

72 See *Preston Jones* v *Preston Jones*, n 68 above. There was in fact some disagreement about whether the criminal standard of proof was required; see *Re L* [1968] P 119, 153–154, 171–172.

73 At 763.

74 s 140 of the Evidence Act 1995 (no 2 of 1995) reads:

> (1) In a civil proceeding, the court must find the case of a party proved if it is satisfied that the case has been proved on the balance of probabilities. (2) Without limiting the matters the court may take into account in deciding whether it is so satisfied, it is to take into account: (a) the nature of the cause of action or defence; and (b) the nature of the subject matter of the proceeding; and (c) the gravity of the matter alleged.

75 See John Sopinka, Sidney N. Lederman and Alan W. Bryant, *The Law of Evidence in Canada* (Toronto. Butterworths, 1992) 141–148.

76 The third standard of proof is required in cases involving issues such as commitment to a mental institution, loss of citizenship and termination of parental rights. Some states require it for less weighty issues, such as proof of a charge of fraud and rectification of a written contract. See generally William Strong, *McCormick on Evidence* 4th ed (St Paul, Minn: West Publishing Co, 1992) 441–445.

77 The third standard has also been expressed as proof by 'clear, convincing and satisfactory', 'clear, cogent and convincing' and 'clear, unequivocal, satisfactory and convincing' evidence. See *ibid* 441.

© The Modern Law Review Limited 1999

The flexible approach

The flexible approach expounded in *Bater* fits into the decision-theoretic framework developed above: the disutility of an error on a serious issue[78] will often be greater than that of an error on a less serious issue.[79] As we have seen, once the utilities in the equations above are non-symmetrical, the justifications for the 0.5 rule are no longer strong. The way in which the *Bater* standard is expressed, however, makes it seriously deficient. The emphasis is on the seriousness of the allegation, which means that the focus tends to be on $U^{-}(\Pi^{v}|\Delta^{d})$. We have seen, however, that a rigorous decision-theoretic approach to standards of proof requires consideration of three other utilities. At a minimum, it is important also to consider $U^{-}(\Delta^{v}|\Pi^{d})$; if this is as great as $U^{-}(\Pi^{v}|\Delta^{d})$ then the balance of probabilities may well be the appropriate standard of proof. While there are some cases where use of the restricted *Bater* approach may be of little consequence, there are others where it will be: the child abuse cases are a possible example of this.[80]

However, even were the flexible standard of proof to be applied carefully, there would still be a number of criticisms which might be levelled at it. The first is that it is vague. The 0.5 rule has the advantage of clarity: both the parties and the fact-finder know what standard will be applied. But when the standard varies according to the seriousness of the issues, which may depend on the individual judge's utility judgments, things are none too clear. It has been suggested that this has undesirable consequences for litigants, who need to know how much evidence to present in order to prevail.[81] This is not, however, a very convincing argument. Even under the precise 0.5 rule, litigants do not generally say 'I will present only so much evidence as will make my case appear more probable than not'. It is in the nature of litigation that one cannot say when one has enough evidence to make something appear more probable than not until one knows the details of the case presented by the other side. In practice, 'than not' becomes 'than the other side's case'.[82] Unless the plaintiff knows the details of the defendant's evidence (and here I take

78 I am conscious of an ambiguity here. A single case may involve a number of distinct issues, each of which needs to be proved. The conventional understanding is that it is the case as a whole that must be established to the appropriate standard of proof. It would therefore be more correct to say 'the disutility of an error in a case involving a serious issue', but I shall omit this qualification for the sake of simplicity.

79 For a different interpretation of the *Bater* standard, see James Logue, *Projective Probability* (Oxford: Clarendon Press, 1995) 153–154. '[W]hat such comments [as those in *Bater* and *Khawaja*] are groping towards', Logue suggests, 'is the notion of a two-dimensional measure [of probability]: we want $p_1 > 0.5$ and resiliency of p_1 to be high compared to $p_1 - 0.5$ (so that the p_2-distribution is concentrated above 0.5)'. While I have considerable sympathy with the notion of two dimensional probability measures, which are intended to capture the resilience of a probability judgment, or the amount of relevant evidence that it is based on, I do not think that this is a satisfactory explanation of the English case law. A p_2-distribution > 0.5 is simply not enough to protect defendants in cases like *Khawaja* where the disutility of an erroneous verdict against them is substantially greater than the disutility of a verdict against the plaintiff. For that reason, I find it better to interpret the case law as requiring, in Logue's terminology, a higher p_1.

80 Those child abuse cases where a higher standard of proof was held to apply have been criticised on this ground, ie for failing to take into account the consequences for the child of a mistaken verdict for the parents. See J.R. Spencer, 'Evidence in Child Abuse Cases – Too High a Price for Too High a Standard?' (1994) 6 J Child Law 160; Victor Smith, 'Sexual Abuse: Standard of Proof' [1994] Fam Law 626. I discuss the appropriate standard of proof in such cases in more detail below.

81 Pattenden, n 31 above, 230. In *R* v *Hampshire CC, ex p Ellerton* [1985] 1 All ER 599, 608, Slade LJ also opined that: 'any sliding scale ... would lead to intolerable uncertainty in application'.

82 It need not do so, but, for the argument that the comparative nature of proof means that it always should, see Ronald J. Allen, 'A Reconceptualization of Civil Trials' in Peter Tillers and Eric D. Green (eds), *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism* (Dordrecht: Kluwer, 1988).

© The Modern Law Review Limited 1999

'evidence' to include the performance of counsel and witnesses), she cannot judge this with any precision. In general, she will always play safe by presenting as much evidence to prove a point as she can find.

Nevertheless, uncertainty may affect decisions other than the parties' assessments of what it will take to satisfy their burdens of proof. Uncertainty may have implications for settlement. A rational settlement decision will be influenced by the parties' estimations of the probability that the plaintiff will win at trial, and the standard of proof to be applied is obviously relevant here. Again, however, it is unlikely in practice that settlement negotiations do involve anything but the most imprecise calculations of the probability of success.[83] The impact on settlement negotiations of uncertainty about the standard of proof to be applied at trial, then, is likely to be minimal.

There is a third decision which may be affected by the standard of proof, and that is the decision of an appellate court reviewing the trial judge's findings of fact.[84] As we have seen, uncertainty about the civil standard of proof has caused problems in the appellate courts, but this has been due to the question of which test should be applied, rather than to uncertainty about the precise gradation of a flexible standard. Even if the flexible standard were to replace the 0.5 rule in all cases, uncertainty about its precise gradation would be unlikely to be a significant problem on appeal. The Court of Appeal might suspect that the trial judge had not cranked the standard up sufficiently to reflect the seriousness of the issues (perhaps the Court thinks that the trial judge applied a 0.6 standard, whereas it would have shifted the standard up to 0.8). This would presumably be an error of law, but while in theory it might give rise to more appeals and, consequently, more uncertainty for the parties, it would be unlikely to do so in practice. It will seldom be obvious to the Court of Appeal just what standard a trial judge has applied. It might try to work backwards from the weight of evidence to the gradation that that weight would have satisfied, and then find that it could only have satisfied a gradation lower than what it regards as proper. But, handicapped as it is by working from transcripts, the Court of Appeal is extremely unlikely to be able to perform this inferential feat. In any case, there would be little point in its doing so, because it could rehear the appeal as an appeal on an issue of fact, and apply its preferred gradation to the evidence. Although this might possibly lead to appeals from parties hoping that the appeal court will apply a differently graded flexible standard and reach a different decision, the fact is that differing assessments of the weight of evidence will usually be more significant than differing assessments of the required standard, so such cases would most probably still be appeals under the 0.5 rule. In any case, there is already a degree of inertia built into the appeals system by the Court of Appeal's general practice of deferring to the findings of fact of trial judges,[85] which means, once again, that a flexible standard of proof is unlikely to have much impact on the appeals process as it operates in practice.

83 See generally Hazel Genn, *Hard Bargaining: Out of Court Settlement in Personal Injury Actions* (Oxford: Clarendon Press, 1987) esp 97–123, 129–138. Genn notes that uncertainty in fact encourages settlement: in effect, it gives the parties something to bargain over. However, her study does not suggest that uncertainty about the legal rules to be applied is a significant factor in creating uncertainty. Uncertainty about the likelihood of winning is caused more by the unpredictability of evidence and cross-examination, and about what damages are appropriate. See further Neil Rickman, *The Empirical Analysis of Litigation: A Survey of the Economics Literature* LCD Research Series No 2/98 (London: Lord Chancellor's Department, 1998) 10–22.

84 And, as a corollary, the parties' decisions on whether or not to lodge an appeal.

85 See Tapper, n 32 above, 223.

© The Modern Law Review Limited 1999

Uncertainty about when a flexible standard should be applied is, as we have seen, a problem at the moment. But this is not, I think, a problem caused by the flexible standard itself, but rather by the lack of case law on when it should be applied. However, considering when and how to decide that the flexible standard should apply does lead us to a larger question: who is best placed to decide when a higher standard of proof is appropriate and how high should that standard be? To this point, I have been presuming that these decisions will be left, as they are now, to the courts, specifically to trial judges with some guidance – at least as to when a higher standard should be applied – from appellate courts. But it is by no means obvious that this is an appropriate means of making such decisions. Decisions about standards of proof involve judgments about the risk of error that the parties should bear, judgments which our use of decision theory has shown to involve the calculation of the utilities of various outcomes. These are the types of policy decision which are better made by the legislature than by judges.[86] However, whatever the conceptual attractions of a system whereby the legislature lays down the appropriate standard of proof for every form of wrongdoing, such a system is unrealistic.[87] Legislatures rarely lay down standards for the proof of civil and criminal wrongs, even when they use legislation to define those wrongs, a practice which reflects not only the fact that the basic standards – beyond reasonable doubt and the balance of probabilities – were judicial creations in the first place, but also the difficulty of laying down such standards in advance. Given the immense range of factual situations which may form the basis of litigation, courts will inevitably be faced with some unforeseen situations where a modified standard of proof may be appropriate.[88]

Nevertheless, the identification of the choice of standard of proof as a policy decision offers some useful pointers on the way the 'whose utilities' question should be decided. Importantly, it suggests that judicial discretion to set standards of proof should be limited, and that the courts should defer to standards set by parliament where these do exist. What is unsettling about the decision to apply the flexible civil standard in *Serio* v *Serio*[89] is that the issue – legitimacy – to which it was applied in that case did have a standard of proof set for it by legislation. Section 26 of the Family Law Reform Act 1969 set the standard for rebutting the presumption of legitimacy as 'evidence which shows that it is more probable than not that the person is illegitimate'. By reducing the standard from proof beyond reasonable doubt, the aim was to make it easier to prove illegitimacy, and it is highly unlikely that Parliament either intended or foresaw that the standard would be changed to 'substantially more probable than not', which was the effect of the decision in *Serio*.[90]

Where the legislature has not set down a standard of proof for an issue, appeal courts should seek to offer clear guidance on the appropriate standard. There is nothing wrong with the adoption of a higher than normal standard of proof in some civil cases, but there are problems with describing that higher standard as one commensurate with the seriousness of the occasion. This leaves the precise gradation to be determined by the individual judge's utilities; inevitably, different

86 See Alex Stein, 'Against "Free Proof"' (1997) 31 Israel L Rev 573, 582–583.

87 On the practical problems of upholding a strict policy/principle distinction, see John Bell, *Policy Arguments in Judicial Decisions* (Oxford: Clarendon Press, 1983) 221–225.

88 An example is *Morrison* v *Jenkins* (1949) 80 CLR 626, a 'whose baby' case.

89 n 71 above.

90 A point made by Lord Lloyd in *Re H*, n 49 above, 8, who noted that 'the legislature has ordained that the presumption of legitimacy can be rebutted on the simple balance of probabilities' and that in *Serio* the Court of Appeal 'read words into the statute that are not there'.

© The Modern Law Review Limited 1999

standards will be applied in similar cases, a violation of the second dimension of the equality principle.[91] Perhaps we could think of judges applying social, rather than personal, utilities to their decisions; it is not obvious, however, that judges are well placed to do this.[92]

This examination of the flexible *Bater* standard leads to its qualified rejection. Initially, there is nothing suspect about a flexible standard of proof, so long as all of the relevant utilities are considered when it is applied. Although frequent use of the flexible standard would lead to more uncertainty in the trial process, the practical effects of this uncertainty would not be especially pernicious. The objections to the flexible standard are objections of principle: it ties the risk of error to be borne by the parties to the utility functions of individual judges, when this risk should ideally be determined democratically, and should not vary between like cases. Having found the flexible standard of proof wanting, let us turn to the other common approach to the proof of serious issues in civil trials.

The prior probability approach

Again, at first glance there is nothing inherently wrong with the prior probability approach. If, as some would have it, adjudicative fact-finding involves ampliative induction, the fact-finder must start by according a small prior probability to the plaintiff's case. This probability will increase on consideration of evidence which is probable given the truth of that case but improbable otherwise.[93] Alternatively, if fact-finding involves explanation based decision-making, the prior plausibility of explanations of the evidence must be taken into account.[94] Beyond the fact that it is not blatantly illogical, however, there are a number of problems with the prior probability approach.

A preliminary point relates to the way in which this approach to the civil standard of proof is sometimes expressed. It is often suggested that serious allegations are harder to prove than less serious ones because serious wrongs are less likely to have occurred. Such a claim is difficult to defend; we simply do not know very much about the general prevalence of fraud, negligence, arson or child abuse, and this makes the claim that serious forms of behaviour are rarer than less serious ones problematic. Even if there is a general correlation between gravity and frequency of wrongdoing, it beggars belief to presume that the correlation always

91 Another reason why leaving the choice of the precise standard of proof to individual judges would be problematic is that there would be a danger that they might be affected by the characteristics of the litigants. This is a particular problem in criminal cases, where there is a justifiable fear that fact-finders will apply a lower standard of proof to defendants who have undesirable characteristics, such as prior convictions. The danger may not be as great in civil cases, but it is still present, and provides an extra justification for setting general standards of proof, even if this involves some loss of expected utility. See Dekay, n 6 above, 114–115.

92 See Bell, n 87 above, 191–194. Contrast the position in criminal trials It may be that just how large a doubt should be in a criminal case before it leads to an acquittal is best determined by considering the relevant utilities; it may be legitimate for the size of the doubt to vary with the seriousness of the case. It is difficult, however, to determine or describe just how large the doubts should be in any particular criminal case, so we pass these policy decisions to a quasi-democratic body: a jury (or, less ideally, a bench of magistrates). A jury is able to symbolise the application of social utilities in a way a judge is not.

93 This is the model associated with Bayesians. See generally David A. Schum, *Evidential Foundations of Probabilistic Reasoning* (New York: John Wiley, 1994).

94 This is commonly referred to as the 'story model'. See Nancy Pennington and Reid Hastie, 'The Story Model for Juror Decision Making' in Reid Hastie (ed), *Inside the Juror: The Psychology of Juror Decision Making* (Cambridge: Cambridge University Press, 1993); Paul Thagard, 'Explanatory Coherence' (1989) 12 *Behavioral and Brain Sciences* 435.

© The Modern Law Review Limited 1999

holds. As Robertson has observed, there may be some minor regulatory crimes which are in fact very rare.[95]

These criticisms might be avoided by modifying the way in which the prior probability approach is phrased. Although, in *Re H and Others*, Lord Nicholls phrased the test in terms of a gravity/frequency correlation, he also expressed the test in a more innocuous manner: 'the inherent probability or improbability of an event is itself a matter to be taken into account when weighing the probabilities and deciding whether, on balance, the event occurred'.[96] The logic of this can hardly be criticised. Instead, the problem is that this is simply a statement of the obvious; if fact-finders need to be reminded that improbable occurrences are less likely to occur than probable ones, the trial process is in deep trouble. If we cannot presume the basic cognitive competence of our fact-finders, then we may be better off tossing coins to decide liability.[97]

In response, one might ask 'where's the harm?'. Perhaps Lord Nicholls did state something that no competent fact-finder needs to be reminded of, but this comment on the civil standard of proof may serve a purpose, not as an admonition to fact-finders, but as reassurance to those who find themselves accused of serious wrongdoing in the civil courts that such accusations will be carefully scrutinised. Perhaps so, but we should be aware that even such seemingly harmless advice may come to distort the fact-finding process. Lord Lloyd was right to say of the prior probability approach that 'there is a danger that the repeated use of [such] words will harden into a formula, which, like other formulae (especially those based on a metaphor) may lead to misunderstanding'.[98] This is, I think, a serious problem, for reasons which are best approached through a point made by Pattenden. Discussing the prior probability approach, she remarks: '[it] can apply equally well to claims which are morally neutral. For example, it may take a lot to convince a judge that a multi-national corporation entered into a contract with a National Health Service hospital for the supply of sterile sutures at 3 am on Christmas morning'.[99] Although Pattenden does not pursue the implications of this example, there are important points to be made about it. It illustrates the indeterminacy of the prior probability approach: just what is it that should be considered improbable? The factual situation? The illegal nature of some conduct? The example also reminds us that even the most innocuous conduct may, in some senses, be improbable. We might think it very improbable that Jones crossed Oxford Street at 4.37 pm on 23 August 1997. But would this really be difficult to prove? Our intuition is that it would not be, but why not? What is the precise relationship here between prior probability, evidence, and proof?

Another example will help us to think these issues through. Consider a lottery with 10,000 tickets. An eyewitness, who is 99.9 per cent reliable, saw the winning ticket drawn and informs us that it was ticket number 642. Should we believe her?[100] The

95 Bernard Robertson, 'Criminal Allegations in Civil Cases' (1991) 107 LQR 194.

96 n 49 above, 17.

97 For discussion of this suggestion, see William Twining, *Rethinking Evidence: Exploratory Essays* (Evanston: Northwestern University Press, 1994, orig 1990) 117–122.

98 In *Re H and Others*, n 49 above, 8–9.

99 Pattenden, n 31 above, 227–228.

100 This problem is known as the 'Todhunter paradox': see I. Todhunter, *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace* (New York: Chelsea, 1949, orig 1865) 400. Certain writers have claimed that conventional probability theory suggests that we should not believe the witness: See L. Jonathan Cohen, 'Can Human Irrationality be Experimentally Demonstrated?' (1981) 4 *Behavioral and Brain Sciences* 317, 329. My analysis draws on Jonathan J. Koehler and Daniel N. Shaviro, 'Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods' (1990) 75 Cornell L Rev 247, 270–271.

© The Modern Law Review Limited 1999

prior probability of that ticket winning is low, but experience of lottery announcements suggests that prior improbability plays no great role in affecting the reliability of reports of winning numbers. To tease out the solution to this paradox, we need to note that, if the eyewitness was indeed mistaken, it is in fact very unlikely that she would have reported ticket 642 as having won, because there were 9,998 other mistakes she could have made. In effect, the improbability of her reporting 642 given that 642 did not win cancels out the improbability of 642 having been drawn, leaving us with a posterior probability of 99.9 per cent that 642 won.[101] The lottery example is useful, because, as we know how many mistakes were possible, it is easy to model the probability of the report of 642 given that some other ticket won. Ordinarily, the probability space for possible mistakes is not so easy to construct, but the key insight to be drawn from the lottery example remains, by and large, valid: events which are very unlikely to happen are also unlikely to be alleged if they did not happen. So when an eyewitness reports having seen Jones crossing Oxford Street, a relevant consideration is: if she was mistaken, why did she report Jones? Why not Smith, or Palmer, or Andrews? If there is no obvious reason for her choosing Jones as opposed to anyone else in the population, the prior improbability of the event plays little role in determining our final judgment of its likelihood. The insight is only by and large valid because some reports of improbable events are in fact quite likely to be made even when they are untrue. In my books, the prior probability of aliens visiting earth and inserting probes up people's rectums is pretty low, but when my neighbour tells me that this is just what happened to him last night I do not jump to the conclusion that it probably did happen; the story has such cultural availability that it is rather likely to be reported when untrue.[102] But if one of my colleagues tells me that she saw a flying saucer on her way into work this morning, and I know that she is as sceptical of such things as I am, that she is aware that her journey lies under a major flight path and that she is not prone to practical jokes, I might credit her story, because I would judge her unlikely to make the report if what she saw was not a flying saucer.

The foregoing analysis suggests that prior probability does not play a determining role in all posterior probability judgments. What is just as important for the fact-finder to think about is the likelihood of the allegation being made, and of the case coming to court, if it is untrue. For example, if allegations of child abuse are very unlikely to be made if untrue, then, even if child abuse is incredibly rare, an allegation of child abuse may make child abuse quite probable.[103] Often,

101 This can be shown formally using likelihood ratios to apply Bayes' theorem as follows:

Odds (642 won given that 642 reported) =

$$\text{Prior odds that 642 won} \times \frac{\text{P(642 reported given that 642 won)}}{\text{P(642 reported given that 642 did not win)}}$$

Substituting figures:

$$\text{O(642 won)} = \frac{1}{9999} \times \frac{0.999}{0.0001} = \frac{0\,999}{0.001} = 999 \text{ to } 1$$

This analysis involves an important simplifying assumption. that all mistakes are equiprobable. This is unlikely to be the case – the most likely error is one that mistakes just one digit of the winning number.

102 See Elaine Showalter, *Hystories: Hysterical Epidemics and Modern Culture* (London: Picador, 1997) 6, 189–201.

103 The analysis presented here sheds some light on a point made by Keating, criticising the House of Lords' decision in *Re H* [1996] 1 All ER 1 (Heather Keating, 'Shifting Standards in the House of Lords' (1996) 8 Child & Fam L Q 157, 160; see also Caroline Keenan, 'Finding that a Child is at Risk from Sexual Abuse' (1997) 60 MLR 857, 864). Keating suggests that the assumption that serious

© The Modern Law Review Limited 1999

we grasp the prior probability/probability of untrue allegation relationship intuitively, which is why we believe lottery results and eyewitness reports of the type 'I saw Jones crossing Oxford Street'. The danger of fact-finders reminding themselves, as the House of Lords suggests they do, that parents are unlikely to have abused their children, is that this intuitive understanding may be destroyed while being replaced with only half of what is required for a rigorous analysis, viz, *both* a prior probability estimate *and* an estimate of the probability of a report given that abuse did not occur.

A single third standard of proof

This evaluation of judicial approaches to the civil standard of proof suggests that, in *Re H*, the House of Lords got it wrong. As an instruction on the standard of proof, the prior probability approach makes little sense. What is more, it has the potential to distort the fact-finding process. If the courts are to modify their approach to proof in serious civil cases, it makes far more sense to raise the standard of proof in order to reflect the importance of the interests at stake. But this still leaves unanswered questions: what sort of cases might require a higher standard of proof? And, given the problems of principle which afflict the flexible standard of proof, how exactly should the standard of proof for serious cases be formulated?

The second question is easy to answer. The flexible standard is flawed because it allows individual judges to make decisions about the weight that should be attached to the rights of litigants in individual cases, and because it allows too much variation in the standard of proof. I suggested earlier that, because of the importance of the equality principle and the general desirability of minimising expected errors, the 0.5 rule should be the default choice in civil cases, and that it should be protected by a degree of inertia. In other words, the courts should only depart from it where the reasons for doing so are strong. Taken together with the problems of inconsistency of application that have afflicted the flexible standard in practice, all this suggests that the flexible standard should be replaced by a single higher standard of proof. The obvious candidate is the American 'third standard': proof by clear and convincing evidence.[104]

The question of when this standard should be applied does not admit of such a simple answer. It will be easiest to respond to it if we keep some fairly clear examples in mind. In the final section of this article, therefore, I review three situations where the third standard might be appropriate. I examine the factors that might determine whether the third standard should apply and suggest whether it is warranted.

---

events have a lower prior probability may be negated by the fact that serious allegations are less likely to be made if untrue. It should by now be clear that things are rather more complicated than this: the availability of claims of child sexual abuse *may* mean that they are often made when untrue. Further, if the sexual abuse of children is rare, it is quite possible for false positives (untrue allegations) to outnumber true positives, even when allegations are not lightly made. A similar problem arises with HIV testing. On the 'looping effect' of the concept of child abuse, see Ian Hacking, *Rewriting the Soul: Multiple Personality and the Sciences of Memory* (Princeton: Princeton University Press, 1995) 55–68. Hacking also discusses the problem of false positives in disease testing at 109–110, the example is a common one in the probability literature. See, for example, Dawid, n 2 above, 432–434.

104 The formulation of the American third standard is not necessarily ideal, as it suggests a standard based on the amount of evidence presented rather than on the fact-finder's degree of belief. It might be preferable to phrase the third standard as 'substantially more probable than not', or 'highly probable'. The latter is advocated in McBaine, 'Burden of Proof: Degrees of Belief' (1944) 32 Calif L Rev 242, 262–263.

© The Modern Law Review Limited 1999

Choosing an appropriate standard of proof: three examples

Some preliminary comments are in order. Civil litigation is not like criminal litigation. In a criminal case, the defendant faces the state which almost invariably has superior resources. In some civil cases, to be sure, the playing field is not level because one party is able to marshal superior resources. But this is not a regular feature of civil litigation (although it may be common in certain types of civil case, which might justify a departure from the norm). When we set a standard of proof, it is the regular case we should have in mind. What is more, even in those civil cases that do pit David against Goliath, the party with superior resources does not have the coercive powers enjoyed by the state in criminal cases: she cannot, for example, arrest and question her opponent. On top of this asymmetry in resources lies an even more important difference between civil and criminal trials: the outcome of a criminal trial is far more serious than the outcome of a civil trial. By this I do not just mean that a conviction in a criminal court can lead to loss of liberty, although it obviously can. Imprisonment accounts for only a small proportion of disposals of convicted defendants.[105] The most common punishment is a fine, and the level of fine imposed is often fairly low; in fact, in strict material terms the loss a criminal defendant faces may often be less than that faced by a civil defendant.[106] The point I wish to make is that a criminal conviction is far more serious in *non-material* terms than is a finding of civil liability. A finding of criminal responsibility has an important censuring function which has no equivalent in civil litigation.[107] Moreover, the stigmatising effects of criminal conviction last well beyond the time of conviction.[108] A finding of civil liability will only rarely have such long term effects.

It is important to bear these broad distinctions in mind when we consider the examples that follow. Another distinction which should be made at this point is that between serious allegations and serious outcomes. We have seen that in criminal cases, whether an allegation is serious or not, the potential outcome (an ascription of criminal responsibility with its stigmatising effects) always is. In a civil case, even when the allegation is serious the potential outcome does not have those serious effects that mark out criminal litigation. In this section, I proceed on the assumption – which I will defend in the discussion to follow – that a higher than normal standard of proof is only required where the outcome (rather than the allegation) is serious.

---

105 Around 5 per cent. See Gordon C. Barclay (ed), *Digest 3: Information on the Criminal Justice System in England and Wales* (London. Home Office Research and Statistics Department, 1995) 37.

106 As Birks remarks: '[p]eople are ruined financially and professionally in the daily grind of civil jurisdiction. Compensatory damages easily run into hundreds of thousands of pounds ...'. P.B.H. Birks, 'Civil Wrongs: A New World' in *Butterworths Lectures 1990–91* (London: Butterworths, 1992) 80.

107 A point made by commentators adopting a variety of theoretical perspectives. See Andrew von Hirsch, *Censure and Sanctions* (Oxford: Clarendon Press, 1993) 6–19, 24–27; R.A. Duff, *Trials and Punishments* (Cambridge: Cambridge University Press, 1986) 39–73, 99–143, 146–148; John Braithwaite and Philip Pettit, *Not Just Deserts: A Republican Theory of Criminal Justice* (Oxford: Clarendon Press, 1990) 88–90, 120–124.

In civil litigation punitive damages arguably have a censuring function. For the contention that this should entail the imposition of a higher standard of proof, see Peter Jaffey, 'Restitutionary Damages and Disgorgement' (1995) 2 Restitution LR 30, 38–40; cf Birks, n 106 above, 79–80.

108 See J.P. Martin and D. Webster, *The Social Consequences of Conviction* (London: Heinemann, 1971); Andrew von Hirsch and Martin Wasik, 'Civil Disqualifications Attending Conviction: A Suggested Conceptual Framework' (1997) 56 CLJ 599; Nigel Walker, *Punishment, Danger and Stigma: The Morality of Criminal Justice* (Oxford: Blackwell, 1980) 142–163.

© The Modern Law Review Limited 1999

Deportation and loss of liberty

On occasion, civil courts deal with cases which potentially have very serious outcomes for the individuals concerned. An example is *R* v *Home Secretary, ex parte Khawaja*,[109] where the applicants had been detained pending deportation on the grounds that they were illegal entrants. They applied for judicial review to challenge the decisions to detain them. The serious outcome was obviously the applicants' loss of liberty; what is more, losing the appeal would effectively mean that they would be summarily deported.[110] the House of Lords decided that the executive bore the burden of proving that the applicants were illegal entrants; the standard of proof was held to be the balance of probabilities, but several of their Lordships expressed the view that this required what I have referred to as the flexible standard.[111]

Although cases such as this do not involve the stigma or censure of a criminal conviction, the seriousness of being deprived of one's liberty and of being deported from a country which, if the immigration officials are wrong, one has a right to be in, suggest that this is a case where a higher standard of proof is appropriate. We can check this intuition against the relevant utilities. We do not have to assign precise figures to the utilities in the equations given above; it will be enough to rank them in order of preference. Taking the person who is challenging the immigration officer's decision to be the plaintiff, $U(\Pi^v|\Pi^d)$ means that a person who should not lose her liberty is set free; $U^-(\Pi^v|\Delta^d)$: a person who should lose her liberty is set free, with the probable consequence that the state will not succeed in deporting her even though she has no right to remain in the country; $U(\Delta^v|\Delta^d)$: the state succeeds in taking away the liberty of someone who has no right to remain free; $U^-(\Delta^v|\Pi^d)$: a person who has the right to be free loses her liberty. A plausible ranking is as follows: $U(\Pi^v|\Pi^d) > U(\Delta^v|\Delta^d) > U^-(\Pi^v|\Delta^d) > U^-(\Delta^v|\Pi^d)$. Under any such assignment, $EU(\Pi^v) = EU(\Delta^v)$ if and only if $P(\Delta^d) > 0.5$, so a higher standard of proof is called for. If we take the difference in utility between the plaintiff's verdicts and the defendant's verdicts to be substantial (which again seems plausible), then the higher standard of proof (by clear and convincing evidence) is clearly justified. The argument for the higher standard is all the stronger when we observe that in cases such as this the state will, as a rule, have substantially greater resources than the immigrant.

There is another situation where a plaintiff may lose her liberty in non-criminal proceedings: proceedings for civil contempt of court.[112] In such proceedings, the criminal standard of proof is required.[113] This has been justified in the following terms: '[a] contempt of court is an offence of a criminal character. A man may be

109 n 57 above.

110 The cases also involved a serious allegation, because the applicants were said to have deliberately misled immigration officials when seeking entry to the country. However, the possibility of loss of liberty would probably have been held to call for a higher standard of proof by itself.

111 Later cases have held *Khawaja* to require proof by a high degree of probability in similar factual situations. See *R* v *Secretary of State for the Home Department, ex p Rahman* [1997] 1 All ER 796; David Jackson, *Immigration: Law and Practice* (London. Sweet and Maxwell, 1996) 850–854. The United States Supreme Court has also held that a higher standard of proof is required in deportation proceedings: see *Woodby* v *Immigration Service* 385 US 276 (1966)

112 Although it is habitual to distinguish between civil and criminal contempt of court, the difference between the two may be more apparent than real. Nevertheless, I would suggest that neither is a truly criminal type of proceeding. See generally C J. Miller, *Contempt of Court* 2nd ed (Oxford: Clarendon Press, 1989) 2–3, 25–45.

113 *Yianni* v *Yianni* [1966] 1 WLR 120; *In re Bramblevale Ltd* [1970] 1 Ch 128; *Dean* v *Dean* [1987] 1 FLR 517.

© The Modern Law Review Limited 1999

sent to prison for it'.[114] The potential loss of liberty certainly justifies a higher standard of proof, but it is loose talk to equate this with the offence being of a criminal character. An adverse finding of fact will quite simply not have the censuring and stigmatising effects of a criminal conviction; nor is committal to prison for civil contempt punitive in the conventional sense: its purposes are said to be remedial or coercive.[115] I would suggest that there is no need for proof beyond reasonable doubt in contempt of court proceedings; proof by clear and convincing evidence will provide sufficient protection for those accused of being in contempt of court, while recognising that the threat of committal for contempt may be the only effective means a successful plaintiff has of having a judgment enforced. What is more, the use of the criminal standard threatens to blur the important distinction between criminal and non-criminal wrongs.[116]

Child abuse

The sexual or physical abuse of a child is a criminal offence. There are, however, many situations where the question of whether a child has been abused arises in civil proceedings. We need a clear example in order to explore the various considerations affecting the appropriate standard of proof, so let us take facts similar to those in *Re H and others (minors)*. A local authority suspects that a child is being sexually abused by her parents. The authority applies for a care order under section 31(2) of the Children Act 1989, which provides that: 'A court may only make a care order ... if it is satisfied ... that the child concerned is suffering, or is likely to suffer, significant harm; and ... that the harm, or likelihood of harm, is attributable to ... the care given to the child or likely to be given to him if the order were not made'. Essentially, then, the local authority must prove that the conduct of the parents means that the child is suffering, or is likely to suffer significant harm. What standard of proof should the local authority have to prove this to? This is not as clear cut as the *Khawaja* scenario. There is a serious allegation (child abuse) and two potentially serious outcomes (a child is left with parents who are abusing it; a child is taken away from parents who are not abusing it).

In the English case law that discusses the standard of proof in child abuse cases, one finds little careful analysis of the interests at stake and how they relate to the choice of standard of proof. For that reason, it may be helpful to approach the issues via a decision of the Supreme Court of the United States. In *Santosky* v *Kramer*[117] the Court held, by a majority, that at a hearing to terminate parental rights on the grounds of permanent neglect, proof by clear and convincing evidence was mandated by the due process clause of the Fourteenth Amendment. The case involved an appeal from a New York family court which had applied the normal civil standard of proof. As the majority pointed out, this standard of proof meant that the parties (the State and the parents) shared the risk of fact-finding error in roughly equal fashion. This, it held, was objectionable, owing to the

114 Lord Denning MR at [1970] 1 Ch 137. In *Dean* v *Dean*, n 113 above at 521, Dillon LJ said that '[i]t has long been recognised that the procedure in contempt is of a criminal nature', but this simply begs the question.

115 See Miller, n 112 above, 25.

116 A similar argument would, I think, hold for the standard of proof in criminal contempt proceedings. Although the sanction for criminal contempt is sometimes said to be punitive, it is still qualitatively different from the criminal sanction. The purpose of punishment for contempt, it has been said, is 'to vindicate the authority of the court' (per Lamar J in *Gompers* v *Bucks Stove and Range Co* 221 US 418 (1911) 441) rather than to express social condemnation.

117 455 US 745 (1982).

 © The Modern Law Review Limited 1999

asymmetry in the interests at stake: on the one hand, the parents' profound interest in retaining custody of their children; on the other, the State's *parens patriae* interest in promoting the welfare of the child. The latter interest, the majority held, was coterminous with the parents' interest in retaining custody, at least until permanent neglect was proved.[118] The minority, however, challenged this last assumption. The children, it argued, only had an interest in staying with their parents if they would provide a good home; to the extent that it was probable that the parents were neglectful, the children had an interest in having links with their parents severed.[119] Seen in this light, the ordinary civil standard of proof was appropriate: the interests on either side were equally weighty, so it was fair to equalise the risk of error borne by each.

What can be learned from *Santosky* is that the appropriate standard of proof in a case like this depends, in large part, on the way in which the dispute is classified. Should it be seen as a case of parents versus state, or of parents versus child? Returning to the situation in a case like *Re H, Santosky* suggests that we need to delineate carefully the interests involved in a decision under section 31(2). The first point to make is that a finding that the parents are causing the child to suffer significant harm under that section does not automatically lead to the imposition of a care order. If the criteria are met, a court should consider whether making one would be in the child's best interests.[120] Because of this, it might be argued that the child's interests (and therefore the local authority's interests, which are parasitic on the child's interests) are not relevant to the factual determination under section 31(2). If so, there would be a very strong argument for following *Santosky* and imposing a high standard of proof on the local authority.

It is not obvious, however, that the two decision-making stages can be so neatly separated. First of all, because the threshold criteria require proof of 'significant harm', it is unlikely that there will be many cases where the criteria are proved but where a care order is not considered to be in the child's best interest.[121] To all intents and purposes, then, the fact-finding under section 31(2) will be determinative of the whole question; the child's interest will not play a significant role at the second stage. The implication of this is that the child's interests should be given weight in the determination of the standard of proof for the section 31(2) question. But this does not tell us what the child's interests are. Might there be, as the majority in *Santosky* found, an identity of interest between parent and child? This just does not seem credible. The child's interests are in remaining with the parents if the parents are not abusive, but not if the parents are abusive. But whether the parents are abusive is just what the court is trying to find out, so that, in a manner of speaking, the court is also trying to find out where the child's interests lie. If there is evidence that the parents are abusive, then there is evidence that the child's interests do not lie in remaining with the parents: this is so no matter how strong the evidence. To hold, as the *Santosky* majority did, that divergence of parent/child interests only eventuates when a finding of neglect is made is to fetishise fact-finding when all we have is evidence and probability.

The points made in the preceding paragraph do not necessarily mean that the risk of error should be spread equally between parent and child. To determine the

118 *Ibid* 615.

119 See especially *ibid* 628 n 13.

120 See *Humberside County Council* v *B* [1993] 1 FLR 257 which held that once the threshold criteria have been established, a court must consider the factors in ss 1(1), 1(3) and 1(5) of the Act.

121 It appears that a care order is most likely to be refused where the local authority has not made clear plans for the child's future. See Richard White, Paul Carr and Nigel Lowe, *The Children Act in Practice* 2nd ed (London: Butterworths, 1995) 194–196.

© The Modern Law Review Limited 1999

appropriate standard of proof, we need to consider the relevant utilities. If we keep the *Re H* situation in mind, and treat the parents as the plaintiffs, what might a plausible ranking be? Consider $U(\Pi^v|\Pi^d)$: no care order is made when the state has no right to intervene in the parent child relationship. $U^-(\Pi^v|\Delta^d)$: no care order can be made when the state has a right to intervene and the child deserves protection. But note that this does not mean that the child will never receive protection. The authorities have been alerted to the possibility of abuse; if it continues, and there is better evidence of it, a future application for a care order may succeed. On the other hand, the abuse may stop. $U(\Delta^v|\Delta^d)$: a care order will almost certainly be made when the child deserves protection. A care order does not, however, sever all links between parents and child; effectively, parental responsibility will be shared between the parents and the local authority.[122] It is possible that the child will be restored to the parents' care at some future date. However, if the abuse was serious this is less likely to happen. We should also note that the care order will not necessarily protect the child from future abuse; she might be abused while in local authority accommodation, or by foster or adoptive parents. $U^-(\Delta^v|\Pi^d)$: a care order is made when the parents have a right to retain care of the child. Again, note that loss of the child might not be permanent, although the more serious the abuse the more likely that it will be. The child may be at more risk than she was in before, and the parents may face considerable social stigma. It is not obvious, however, how significant this last consideration should be: parents are protected from adverse publicity by the anonymity of care proceedings. Although the removal of their child may become known in their community, the reason for it might not.

What is obvious is that there is a large degree of uncertainty about the effects of the possible verdicts in care proceedings, and this does not allow one to have much confidence in judgments about the appropriate standard of proof. A case can be made for ranking the utilities as follows: $U(\Pi^v|\Pi^d) > U(\Delta^v|\Delta^d) > U^-(\Pi^v|\Delta^d) > U^-(\Delta^v|\Pi^d)$. As in deportation cases, this would justify a higher standard of proof.[123] However, given the uncertainty involved in ranking the utilities and the fact that the difference in utility between the defendant's and plaintiff's verdicts does not appear to be large, the default standard – proof on the balance of probabilities – seems most appropriate.

Proof of crime in civil proceedings

It is not clear what standard of proof is required when criminal conduct is alleged in a civil case. There is some authority for saying that in such cases the standard of proof should be the criminal standard. A number of older cases support this view,[124] which has been approved by the Privy Council.[125] However, the Court of Appeal and the House of Lords have both held *Hornal* v *Neuberger Products Ltd*[126] to be authority for the proposition that the civil standard (albeit the flexible

122 See Children Act 1989, s 33(3).

123 But cf Mary Hayes, 'Reconciling Protection of Children with Justice for Parents in Cases of Alleged Child Abuse' (1997) 17 LS 1, 15–17.

124 A useful discussion is R.M. Stonham, 'Proof of Crime in Civil Proceedings' (1934) 8 Aust LJ 207.

125 In *The People of the State of New York* v *Heirs of the Late John M Phillips and Others* [1939] 3 All ER 952 the Privy Council heard an appeal involving allegations of conspiracy. The trial judge had held that the plaintiffs had to prove their case as clearly as in criminal proceedings, and Atkin LJ remarked, at 955, that '[t]heir Lordships consider this not to be ill-founded. The proposition of the judge has been laid down time and again in the courts of this country and it appears to be just and in accordance with the law'.

126 n 40 above.

© The Modern Law Review Limited 1999

one) is appropriate.[127] In at least one recent case, a civil court has applied the criminal standard of proof to an especially serious case (one involving an allegation of murder).[128] All of these decisions should be contrasted with the judgment of the House of Lords in *Re H* which, to all intents and purposes, decided that allegations of one type of conduct (child abuse), which in many situations could give rise to criminal proceedings, need only be proved on the basic balance of probabilities.

Are there good reasons for applying a higher standard of proof in civil cases involving allegations of criminal conduct? I am not persuaded that there are. At the beginning of this section, we noted the significant differences between criminal and civil proceedings; it is important to keep this distinction in mind as we examine the problem of criminal allegations in civil cases. A finding of civil liability does not censure the defendant, nor will it have many of the collateral effects of a criminal conviction. Consequently, where, as in *Halford* v *Brookes*,[129] an allegation of murder is made in a civil case, the outcome is not serious even though the allegation, in some senses, may be. In fact, on consideration we may decide that the distinction between serious outcomes and serious allegations is none too clear. What makes an allegation serious is very often the fact that it has a potentially serious outcome. In a criminal case, the prosecution's allegations are serious because they may lead to punishment; an allegation of crime in a civil case never has this added dimension of seriousness.

In response to the suggestion that the civil courts judge allegations of criminal conduct by the simple balance of probabilities, it might be argued that a finding of theft, fraud or murder will stigmatise the defendant and that, therefore, a higher standard of proof is appropriate. There is some truth in this, especially where the allegation is of murder or rape.[130] However, we should remind ourselves once more that this sort of stigma is different from the censure imposed by a finding of guilt in a criminal court; we should also note that there is a degree of circularity at work here. The more we treat allegations of crime made in the context of civil litigation as serious, and signify this seriousness by adopting a higher standard of proof, the more they become serious.[131] When the civil courts require the satisfaction of a high standard of proof before finding a defendant liable for murder, they help to justify the treatment of the defendant as a murderer. At the risk of stretching the

127 *Nishina Trading Co Ltd* v *Chiyoda Fire and Marine Insurance Co Ltd* [1969] 2 QB 449, 462, 464 (CA); *R* v *Home Secretary, ex p Khawaja* [1984] 1 AC 74, 112 (HL).

128 *Halford* v *Brookes and Another, The Times* 3 October 1991. Rougier J, however, remarked that 'the distinction ... [between the criminal standard and the *Bater* standard] is, for present purposes, really one without a difference'. Contrast *Miles* v *Cain, The Times* 15 December 1989 where the Court of Appeal held that the *Bater* standard was appropriate in a case involving an allegation of rape, and *Francisco* v *Diedrick, The Times* 3 April 1998 where, following *Re H*, the prior probability approach was taken; cf also *Halford* v *Brookes* [1991] 1 WLR 428, where, during an appeal on whether the action in tort was barred by the limitation period, Lord Donaldson MR remarked, at 447: 'there is no necessary inconsistency between pursuing the claim against the second defendant and his acquittal on a charge of murder. Quite apart from the fact that the standard of proof required is different, a lower standard being required in civil proceedings ...'.

129 n 128 above.

130 See, for example, *The Times* on 26 March 1998 reporting that Tony Diedrick, whom the High Court had found liable for the killing of Joan Francisco, had gone into hiding. The trial was reported in the newspapers in the preceding days.

131 The civil courts have contributed to undermining the distinction between civil and criminal proceedings by using the heavily charged word 'guilty' rather than the morally neutral, and technically correct, 'liable' or some other neutral term. See, for example, Rougier J in *Halford* v *Brookes*, n 128 above ('guilty of murder'), Denning LJ in *Hornal* v *Neuberger Products*, n 40 above at 973 ('guilty of fraud'); Lord Oaksey in *Preston Jones* v *Preston Jones* n 68 above, 408 ('guilty of adultery').

© The Modern Law Review Limited 1999

English language to breaking point, it might be argued that, because civil courts lack many of the safeguards associated with serious criminal proceedings – such as trial by jury, the stricter rule against hearsay and the need to prove *mens rea* – they simply cannot meaningfully find that a 'crime' has been committed, even if they do impose the criminal standard of proof. Perhaps we should go so far as to replace the phrase used as the heading of this subsection with 'proof of conduct which, if proved in criminal proceedings, would be a crime'.

There are no good reasons for applying a higher standard of proof when allegations of crime are made in civil proceedings, and there are good reasons for not doing so. $U^{-}(\Delta^{v}|\Pi^{d})$ will, as a rule, only have a greater negative value than $U^{-}(\Pi^{v}|\Delta^{d})$ if we are moved by pleas of the negative consequences of stigma for the defendant. The same goes for $U(\Delta^{v}|\Delta^{d})$ being greater than $U(\Pi^{v}|\Pi^{d})$. Further, if we do impose a higher standard of proof in such cases, it might be difficult to justify the violation of the second dimension of the equality principle which would result: plaintiffs who had suffered harm as a result of conduct which happens to be criminal rather than merely tortious would find it harder to gain redress than would other plaintiffs.

Conclusion

In this article, I have developed a conceptual framework for understanding standards of proof in civil litigation, and defended the 0.5 'more probable than not' standard as the default standard for civil cases. I have argued that the principles developed in the case law are confused and difficult to defend, and that the law would be improved by the adoption of a single third standard of proof. To pull the various strands of my argument together I used three examples to demonstrate when and how the application of the higher standard of proof might be justified.

The arguments that I have presented are, of course, open to the criticism that they attempt to impose precision and clarity on an area where these are not virtues. It is certainly true that there is little uncontested understanding of just what it means to prove a case to a certain standard and that, while fact-finders may know when the cases they judge have been proved, they may well find it difficult to articulate precise reasons in every case. It is not surprising that the subject of the civil standard of proof moved Morris LJ to trot out a version of Holmes' well known aphorism: 'the life of the law is not logic but experience'.[132] However, while I am very much aware of the shortcomings of attempts to explain proof in terms of decision theory and probability theory, I am not convinced that the project of providing a clear conceptual framework for understanding the civil standard of proof is misconceived.[133] Whatever the nature of the *process* of proof in legal fact-

132 In *Hornal* v *Neuberger Products*, n 40 above, 978.

133 Given that I have elsewhere expressed some scepticism about the ability of conventional probability theory (CPT) to capture all aspects of forensic proof (see my 'Bayesianism and Proof' in Michael Freeman and Helen Reece (eds), *Science in the Courts* (Aldershot: Dartmouth, (1998)), I should perhaps expand on this point, because I have opened myself up to accusations of inconsistency in using CPT to analyse standards of proof in this article Whatever is the correct theory of forensic proof, it seems inconceivable that the theory would not (a) accept that standards of proof vary and (b) explain this through a concept analogous to the varying utilities of verdicts in different situations. Writers who have developed theories of proof which are not based on CPT seem to accept this implicitly. See L. Jonathan Cohen, *The Probable and the Provable* (Oxford: Clarendon Press, 1977) 255–256; Ronald J. Allen, 'The Nature of Juridical Proof' (1991) 13 Cardozo L Rev 373, 413; Stein, n 26 above, 323–327, 333–338 My own view is that CPT can be a fairly – though not completely – accurate description of forensic proof. If that is correct, the decision-theoretic analysis provides a

© The Modern Law Review Limited 1999

finding, we need some basis for understanding why we have *standards* of proof and why we may wish to vary them in different adjudicative contexts. The price English law has paid for ignoring this is clear: a confused and inconsistent case law.

There is, however, a more fundamental criticism that might be made of my argument for conceptual clarity in the law on standards of proof. Even if one accepts that a clear and coherent understanding of standards of proof is attainable, one might still conclude that it would not be desirable for the courts to use it to develop the law in this area. There are sometimes good reasons for conceptual ambiguity in the common law. By not committing themselves too deeply to any particular theory, courts narrow the scope for disagreement among judges and allow themselves some leeway to respond to the merits of individual cases.[134] There are many areas of the law where this is a telling argument, but the area considered in this article is not, I think, one of them. It is not just that there is ambiguity in the current law governing the civil standard of proof; there is also little coherence to the way the law has developed. When judges in trial courts attempt to apply the current law, this lack of coherence may distort the fact-finding process. The early child abuse cases, and *Re M*[135] in particular, provide examples of the pernicious effects this incoherence may have. Although my analysis in this article has been fairly technical, I am not suggesting that judges should become experts in decision theory; indeed, I have attempted to develop a framework that allows for fairly coarse distinctions to be drawn between cases.

There is a further reason why there is value in bringing some order and tidiness to the law on the civil standard of proof. In this article I have concentrated, by and large, on the conceptual nitty gritty of standards of proof. What I have paid little attention to is the wider, symbolic role that standards of proof play in the legal system. In the jurisprudence of the United States, where the conceptual underpinnings of standards of proof are better understood,[136] so too is this symbolic or expressive role. '[E]ven though the labels used for alternative standards of proof are vague and not a very sure guide to decisionmaking', the Supreme Court has said, 'the choice of the standard for a particular variety of adjudication does ... reflect a very fundamental assessment of the comparative social costs of erroneous factual determinations ... [A] standard of proof represents an attempt to instruct the factfinder concerning the degree of confidence our society thinks he should have in the correctness of factual conclusions for a particular type of adjudication'.[137] This is something that vague talk about a standard of proof which varies with the seriousness of the allegation simply is not able to capture.

---

reasonably robust account of standards of proof. However, the fact that CPT does not map neatly on to forensic proof provides, I think, another justification for protecting the 'more probable than not' standard with a degree of inertia, because one cannot grade judgments of probability and utility sufficiently finely to justify moving to a, say, 0 6 standard of proof.

134 See N.E. Simmonds, 'Bluntness and Bricolage' in Hyman Gross and Ross Harrison (eds), *Jurisprudence: Cambridge Essays* (Oxford: Clarendon Press, 1992).

135 n 67 above.

136 A good example is Harlan J's concurring opinion in *Re Winship* 397 US 358 (1970). One can appreciate the merits of this clarity by considering the opinions in *Santosky* v *Kramer*, n 117 above. Although I think the majority was probably wrong to impose a higher standard of proof in that case, its reasons for doing so were clear, and allowed the minority to pinpoint the areas of disagreement.

137 *Re Winship*, n 136 above, 378–379 (Harlan J) For an analysis of the expressive function of standards of proof, see C.M.A. McCauliff, 'Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?' (1982) 35 Vanderbilt L Rev 1293.

© The Modern Law Review Limited 1999

Theorising Utility Regulation

Tony Prosser*

The Department of Trade and Industry Review of Utility Regulation marks a final recognition that the debate on such regulation is here to stay and will not disappear with the development of further competition or reforms of competition law.[1] Within the European Union a similar debate is taking place with the liberalisation of utility markets, especially telecommunications.[2] Utility regulation has also been controversial worldwide; in the United States, long regarded as its natural home, the relationship between deregulation and new regulatory techniques is now being reconsidered and many of the problems faced earlier in the UK about how to liberalise markets are being faced within a culture which has given a much greater role to identifiably regulatory institutions.[3] Finally, many other countries (for example in Central Europe) are, as the result of their privatisation programmes, having to grapple with the creation of regulatory institutions in cultures not necessarily sympathetic to the concept of relatively independent agencies.

In this article I will not attempt to deal in detail with reform proposals such as those in the Review, important as they are. Instead, I wish to suggest that the theoretical underpinnings implicit in much of the debate are unsatisfactory. In particular, two possible theoretical approaches to regulation are inadequate to bear the weight of understanding, or reforming, regulatory practice. These are a bilateral or contractual model of regulation in which the primary relationship is between regulator and regulated firm; a variant of this can be found in the economist's favourite critical model of 'capture theory'. The second approach (which indeed appears in the Review) is that of stakeholder theory with the regulatory agency seen as the centre of a web of relations in which the firm may be only one interest out of many, and this has something in common with the proceduralism popular in some legal writing. I shall attempt to suggest, using material primarily relating to UK utility regulation, that neither of these approaches can capture the complexity of the regulatory task, either for explanatory or prescriptive purposes. I shall conclude by speculating as to how a more successful approach to studying regulation can be developed by returning to political and constitutional theory and adopting a rights-based approach to regulation.

* School of Law, The University of Glasgow.

I am most grateful for comments on a draft of this article from Cosmo Graham, Douglas Lewis, Colin Scott, John Tasioulas, Charlotte Villiers and the anonymous referees of the MLR A much earlier version was presented as a Scrymgeour Lecture at the University of Dundee on 22 January 1997, and I am grateful to the organisers and those who commented.

1 Department of Trade and Industry, *A Fair Deal for Consumers. Modernising the Framework for Utility Regulation* Cm 3898 (1998) (the Green Paper); *A Fair Deal for Consumers. The Response to Consultation (London: Department of Trade and Industry, 1998).*

2 For an introduction to these debates see C. Scott, *Competition and Coordination: Their Role in the Future of European Community Utilities Regulation* (London: Centre for the Study of Regulated Industries, 1995) and 'Changing Patterns of European Community Utilities Law and Policy. An Institutional Hypothesis' in J. Shaw and G. More (eds), *New Legal Dynamics of European Union* (Oxford: Clarendon Press, 1995). Particularly important questions have arisen with the establishment of new forms of regulatory bodies for telecommunications in France, Germany and Italy.

3 See eg R Pildes and C. Sunstein, 'Reinventing the Regulatory State' (1995) 62 *University of Chicago Law Review* 3.

© The Modern Law Review Limited 1999 (MLR 62:2, March) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

The complexity of regulation[4]

Many claims about regulation, both in practical policy and in economic theory, are misleading; regulation is a much more complex task than either would suggest. For example, in both, regulation has been seen as a temporary way of protecting customers of monopolies and thus 'holds the fort' until markets replace it with the free play of competition; like the state in Marxist theory, regulation then withers away.[5] Moreover, it has been argued that regulation can ideally be accomplished by the application of economic reasoning designed to achieve the single goal of maximising economic efficiency, mainly allocative efficiency.[6] This provides the legitimacy of the regulator through his or her economic expertise; what should be avoided, it is argued, is sullying this with social goals which are for elected governments, not unelected regulators. It does not mean that the regulatory task is necessarily a simple one, for the content of the economic theory applied is often highly controversial and the regulator still needs to ensure an adequate flow of information from the enterprise. Nevertheless, it does provide a potentially coherent set of principles within which the regulatory task can be addressed.

The withering away of the regulatory task and the central rationale of allocative efficiency have had little correspondence with the reality of utility regulation in the UK, in the EU or indeed in the USA. Regulation has shown little sign of withering away; rather it has become more complex as time has passed, as is recognised in the Review in its recognition of an indefinite future for some regulatory controls.[7] This would not in itself matter very much for the view of regulation outlined above in which efficiency maximisation is the predominant regulatory task; it would simply mean that providing market surrogate controls on monopoly would be permanent rather than temporary. However, regulation does not cease with the end of monopoly. One unexpected element in UK utility regulation is the extent to which the regulators have taken on a new task, that of regulation for competition, for example through liberalising the domestic energy markets where the new markets have only been created through constant regulatory pressure in the face of foot-dragging by regulated companies. Moreover, market creation also involves creating a substructure of rules and other institutional and normative devices. Some of these are market-constitutive in the sense that we cannot envisage a functioning market without them; others are designed to create the proverbial level playing field, such as the continuing prohibitions of undue preference or undue discrimination by a dominant firm. Regulation thus has an important role in market creation; and it is not merely temporary, for these rules and other pro-competitive norms need continuing policing.[8] There are advantages in this being undertaken by specialist sectoral regulators which know their firms more intimately than could a general competition authority, especially where the operation of the competition authorities is attended by as many problems as it is in the UK; the reform of competition law currently in hand will not change the desirability of more

4 The arguments in this section are based on research published in much greater detail as T. Prosser, *Law and the Regulators* (Oxford: Clarendon Press, 1997).

5 The most influential statement of this view in relation to UK utility regulation was S. Littlechild, *Regulation of British Telecommunications' Profitability* (London: Department of Trade and Industry, 1984).

6 See C.D. Foster, *Privatization, Public Ownership and the Regulation of Natural Monopoly* (Oxford: Blackwell, 1992) esp ch 9.

7 Green Paper, n 1 above, para 2.8.

8 See C. Shearing, 'A Constitutive Conception of Regulation' in P. Grabosky and J. Braithwaite (eds), *Business Regulation and Australia's Future* (Canberra: Australian Institute of Criminology, 1993).

© The Modern Law Review Limited 1999

specialised, sectoral regulators. As the Review has recognised, '[r]egulatory decisions do not become easier in competitive markets. If anything they become more difficult where more interests have to be weighted'.[9]

In principle once more the growth of regulation for competition does not contradict the view of regulation which insists on its economic rationale in the promotion of allocative efficiency; on this view, regulation for competition serves to promote and police the free markets which are the best way of securing such efficiency. However, across utility regulation it has proved impossible to separate economic approaches to regulation from the broader political and social framework. This can be illustrated vividly from the case of price control. Much of the hard work of the regulators here has been needed because price formulae were set far too generously by government on privatisation in order to permit a successful flotation of shares; this then permitted the recovery of excess profits after sale which in turn created a political outcry and forced regulators to take action both through tough periodic price reviews and through action outside the planned structure of the reviews such as encouragement of profit-sharing even outside the periodic review process.[10] Even apart from the political context, the process of price control has proved to be much less straightforward than originally envisaged. Setting price controls is more an art than a science, and has involved the evaluation of widely different arguments from the various parties involved, a far more procedurally complex process than anticipated.

Finally, it has proved impossible to avoid the intrusion, as some would see it, of social concerns in regulation. If we take regulation other than that of utilities, it is clearly the case that regulation has had an important social dimension. For example, regulation of broadcasting has long been concerned not just with economic goals but with the aim of diversity, ensuring a service which caters for a range of different tastes and interests, and this will not end with the development of channel profusion through digital transmission.[11] In the case of utilities, social regulation has normally, though not exclusively, taken the form of supporting universal service; for example in telecoms, '[t]he concept of universal service is based on the premise that telecommunications services now play such a fundamental role in our society that all people, whoever or wherever they are, must have access to a certain basic level of telecommunications facilities and services if they are to participate fully in modern society'.[12] This has influenced tariff setting through the encouragement of light user tariffs; it has also formed the basis for rules in the liberalised domestic energy markets to prevent 'cream-skimming' which would leave the poor with existing suppliers whilst the rich can benefit from competition and lower prices. Moreover, rather than representing isolated interventions in otherwise competitive markets certain social norms, for example those setting limits to 'cream-skimming', are actually constitutive of the markets themselves through defining their limits. Despite early expectations otherwise, utility services, like broadcasting, cannot be regarded as ordinary commodities but have an essential role in social cohesion

9 Green Paper, n 1 above, para 7.13

10 This perhaps appears most clearly in the case of the water industry, for which see Prosser, n 4 above, 127–132; it also applied to the other utilities and was the basis for the imposition of the 'windfall tax' on their activities by the incoming Labour Government.

11 See D. Goldberg, T. Prosser and S. Verhulst (eds), *Regulating the Changing Media* (Oxford: Clarendon Press, 1998) esp chs 1 and 9.

12 OFTEL, *Universal Telecommunication Services: Proposed Arrangements for Universal Service in the UK from 1997* (London: OFTEL, 1997); this will be discussed more fully towards the end of this article.

 © The Modern Law Review Limited 1999

and this has resulted in further tasks for the regulators, as is once more recognised in the Review.[13]

It is important to state at this point that avoiding giving social responsibilities to regulators appears to be supported by a powerful argument from democracy, an argument that has been made frequently by regulators themselves. This is that, whilst economic matters can be resolved through the regulators' expertise, social concerns involve matters of policy which can only be determined by Parliament through its democratic mandate; as the Director General of Water Services put it, '[s]ocial and sectoral considerations, which involve distributional issues and intervention in market processes, rightly involve politicians'.[14] A similar view appears in the Review; 'the Government, not regulators, should determine the social objectives of regulation' and as a result it proposes the issue of statutory guidance by Government to regulators in each sector.[15] This proposal is highly desirable as a means of providing a clearer framework within which regulators will work. Yet in the end the argument from democracy is unconvincing as a justification for a strict separation between economic and social objectives in the regulators' work. General elections are not determined by the level of British Telecom's light user tariff. At best, elections determine broad differences of approach which may include various assessments of the role of distribution in politics; they do not concern the details of implementation. This does indeed warrant the giving by government of policy guidance in general terms to regulators, though it remains to be seen how willing governments will be to develop hostages to fortune in this way; it should not be forgotten that any measures with significant financial implications (and this would include the most controversial such as energy levies) will not be included in guidance but in new, specific primary or secondary legislation.[16] The other meaning of the argument from democracy is that ministers are accountable to Parliament and so policies can be questioned more effectively and openly when they are made by ministers rather than by regulators. Yet even before the Scott report no one could have seriously dissented from the view that ministerial responsibility in this sense is a concept with minimal application to the real world of government and it neither guarantees the supply of information nor the availability of sanctions. In practice, the regulators have been far more open and accountable in their decision making than have ministers; '[i]n many ways [the Office of Telecommunications] provides a model from which ministerial government departments could learn'.[17]

The utility regulators thus have a variety of different tasks which cannot be reduced to any single logic, economic or otherwise. This then seems to require a pluralist approach to regulation; the regulators' practice, and indeed the legal duties which apply to them, reflect a variety of different rationales.[18] As a result they have come to resemble 'governments in miniature', a phrase originally employed about the early nineteenth-century commissions such as the Poor Law

13 Green Paper, n 1 above, ch 5.

14 I. Byatt, *The Regulatory Scene* (Birmingham: OFWAT, 1996). This concern was also central to the reluctance of the Director General of Gas Supply to become involved in the promotion of energy efficiency schemes; see Environment Committee, *Energy Efficiency: The Role of OFGAS*, HC 328 (1993–94) and Prosser, n 4 above, 110–113.

15 Green Paper, n 1 above, paras 2.13–18, 5.4; *Response to Consultation*, n 1 above, paras 10–14, 57–62.

16 Green Paper, n 1 above, para 2 19; *Response to Consultation*, n 1 above, para 14.

17 C. Scott, C. Hall and C. Hood, 'Regulatory Space and Institutional Reform: The Case of Telecommunications' in Centre for the Study of Regulated Industries, *Regulatory Review 1997* (London: Centre for the Study of Regulated Industries, 1998) 231, 250.

18 For the legal duties see Prosser, n 4 above, ch 1; T. Prosser, 'Privatisation, Regulation and Public Services' (1994) *Juridical Review* 3.

© The Modern Law Review Limited 1999

Commission.[19] This is not of course to maintain that the utility regulators' powers are in practical, or indeed legal, terms, as extensive as those of government departments; in many ways they are severely restricted.[20] Rather it is to suggest that no single logic can or should form a basis for their decision making, and they should not be seen as capable of implementing a mandate of simply applying government-issued guidance. Nor will this be fundamentally changed by the proposal in the Review to replace the regulators' somewhat incoherent current pattern of statutory duties with a single primary duty to exercise their functions in the manner they consider best calculated to protect the interests of consumers, to which other duties would be secondary.[21] There is no simple and universal answer to what is in the interests of consumers (nor indeed a single model of consumer); in many situations maximising competition will clearly be in consumers' interests, but in others this may not be possible, or limits to the free play of market through universal service provision may be more appropriate.

Of course this plurality of regulatory objectives immediately raises the problem of the regulators' democratic legitimacy. If regulators cannot be seen as implementing a mandate laid down by an elected government, one possible answer to this would be to concentrate on the procedures adopted and to see the only suitable reaction to such pluralism as being the development of procedures which ensure that as many different viewpoints as possible are made available to the regulator. Rather than gaining legitimacy through economic expertise, the regulator would do so through wide participation in decision making.[22] I shall return to the potential role and limits of proceduralism towards the end of this article and shall argue that it represents a necessary but not a sufficient step in achieving regulatory legitimacy.

To stress the plural nature of the regulatory task is not enough. What is also necessary is to situate these characteristics of the regulators in more general theories of regulation and in turn to see what implications these theories have for models of how regulation works and how it should be designed. It seems to me that one can identify two radically different theories of the regulatory task, neither of them fully satisfactory in understanding the developments which I have outlined or in providing normative recommendations for reform.

The regulatory contract

The first theory or model is that which underlay the basic design of utility regulation in the UK; it is that of regulation as a form of bilateral contract between regulator and dominant regulated firm, the latter being in practice the enterprise being privatised after having benefited from a legal monopoly. The concept of such a contract seems to have underlain much of the institutional design adopted through creating a model in which the firm is privileged over other interests which may be affected by regulatory decisions.

For example, I mentioned above that many of the regulatory problems arose from the initial price formulae set by government, not by the regulators. These

19 The original source for this phrase is the *Law Magazine*, quoted in H. Arthurs, *'Without the Law': Administrative Justice and Legal Pluralism in Nineteenth-Century England* (Toronto: University of Toronto Press, 1985) 169.

20 For a good statement of this point in relation to telecommunications see Scott, Hall and Hood, n 17 above.

21 Green Paper, n 1 above, paras 3.3–9; *Response to Consultation*, n 1 above, para 16.

22 C. Scott, 'The Proceduralization of Telecommunications Law' (1998) 22 *Telecommunications Policy* 243.

 © The Modern Law Review Limited 1999

were set through bilateral negotiation between government and dominant company; other interests such as consumers were excluded. It is revealing to quote the first Director General of Gas Supply on the process: '[w]hen, in 1986, I asked if I could have a set of working papers on the construction of the formula I was told that certain forecasts had been made by industry experts and it had been, in the final analysis, a judgement call ... when I pursued the issue to try to understand the general thrust of the judgement I was slightly disturbed to hear the value of "X" had been set "to get the company off to a good start" '.[23]

A further example of the privileging of the firm through the adoption of an underlying bilateral contractual model of regulation can be found in the all-important procedures for modifying the licences under which the regulated firms operate; it is these procedures which are used to modify price control formulae and to incorporate new conditions such as those limiting disconnections and, more recently, relating to fair trading. The details vary from case to case, but in outline an amendment can be made if the firm agrees, subject to very limited procedures for publication and the receipt of comments by the regulator when firm proposals are agreed.[24] If the firm does not agree, the regulator can make a reference to the Monopolies and Mergers Commission and will then be in a position to impose a modification taking into account the Commission's report. The latter cumbersome procedure has been used relatively rarely, though its use has increased a little recently, most notably with the recent reference relating to British Gas' transportation and storage charges.[25] The point is that, as a matter of law, the licence can be amended if the firm agrees, irrespective of the views of others affected. Moreover, the 'appeal right' of a Monopolies and Mergers Commission reference, though made at the initiative of the regulator, in practice will only be exercised where the regulated firm will not accept a proposed licence modification and not, for example, at the instance of a consumer group which objects to a change agreed between firm and regulator.[26] It has to be said immediately that the strict legal position is misleading as all the regulators have in practice adopted means to ensure early consultation of consumers and others in the process of licence amendment, with a considerable degree of success in most cases.[27] My point is not that the practice of the regulators has been such as to privilege the firm; rather that the legal basis on which the regulators operate is inadequate to cope with the complexity of this regulatory practice. Nor do the Review proposals change this, concentrating on relatively minor reforms to the Monopolies and Mergers Commission procedure rather than changing the scope of its operation through, for example, creating rights of third-party appeal.[28]

Finally, the stress on regulation as a bilateral relationship between regulator and dominant firm has meant that it is unsuited to the new environment of multi-

23 J. MacKinnon, 'Conference on Regulatory Reform' (1993) 4 *Utilities Law Review* 119, 119–120. For a fuller consideration of the relationship between the privatisation process and regulation see J. Kay and Z. Silberston, 'The New Industrial Policy: Privatisation and Competition' (1984 Spring) *Midland Bank Review* 8.

24 The relevant provisions are the Telecommunications Act 1984 ss 12–15; Gas Act 1986 ss 23–27; Electricity Act 1989 ss 11–14; Water Industry Act 1991 ss 13–16; Railways Act 1993 ss 12–15.

25 Monopolies and Mergers Commission, *BG plc. A Report Under the Gas Act 1986 on the Restriction of Prices for Gas Transportation and Storage Services* (London: Monopolies and Mergers Commission, 1997).

26 For a similar criticism of the licence modification procedures see Scott, Hall and Hood, n 17 above, 241.

27 For a detailed description of the procedures see Prosser, n 4 above, 83–85, 113–115, 144–147, 177–178, 198–199.

28 Green Paper, n 1 above, paras. 7 47–79; *Response to Consultation*, n 1 above, paras. 91–4; cf the new arrangements in general competition law: Competition Act 1998 s47.

© The Modern Law Review Limited 1999

enterprise competition between several firms in parts of the telecoms and energy markets. I suggested earlier that the development of this environment has not meant that regulation has withered away but that a new type of regulation for competition has emerged instead. This is the reason why the Director General of Telecommunications found it necessary to introduce fair trading conditions which proved highly controversial and to make their acceptance by BT a condition of introducing a revised version of the pricing formula.[29] In particular, the patchy coverage of the original 'behaviour specific' licence conditions and the cumbersome procedure for plugging gaps through licence modifications, together with the lack of interdict or damages as remedies for injured competitors, limited the ability of the regulator to respond rapidly; a problem which has been even greater, incidentally, in the bus sector where the ordinary competition authorities were used. It is however questionable whether even after the new conditions an unrevised Telecommunications Act provides an adequate enforcement base for OFTEL to act as a real competition authority.[30] In the case of gas, some limited changes were introduced to prepare for a competitive domestic market by the Gas Act 1995, but these fell far short of a general recasting of the regulator's role to cope with the new environment.[31]

The legal basis for regulation (though not necessarily regulatory practice) was thus created on the assumption that the primary regulatory relationship was with a single firm. Not only did this afford little room for other interests such as consumer groups, it assumed that the sole task of regulators would remain that of policing the actions of a single monopoly provider rather than policing increasingly competitive, multi-enterprise markets. A further step which this model has facilitated is to conceive of the relationship between regulator and dominant firm as a form of bilateral contract, and so to preclude amendments to the regulatory arrangements which increase the burden on firms. Thus this concept of a regulatory contract has been used as a basis for criticism of regulators and government for what has been seen as over-demanding regulation. For example, Cento Veljanovski has claimed that

> there was a feeling that the prospectus would provide a binding commitment or bargain between Government and shareholders ... It was envisaged that [regulation] would operate with a light touch, based on a number of simple and stable rules which would be revised at set intervals ... If one group develops legitimate expectations which are disappointed or there is continuing uncertainty about what are legitimate expectations then it becomes impossible for business to plan with confidence and regulators' actions amount to expropriation of shareholder wealth.[32]

This model has not proved attractive to the regulators themselves; to quote the Director General of Water Services, '[t]here is some uncertainty about what constitutes a regulatory contract. I am not entirely happy with the word. ... Such a

29 OFTEL, *Pricing of Telecommunications Services from 1997: OFTEL's Proposals for Price Control and Fair Trading* (London: OFTEL, 1996). The fair trading conditions successfully survived a challenge by judicial review; see *R* v *Director General of Telecommunications, ex p British Telecommunications plc*, CO 3596/96, QBD.

30 C. Scott, 'Deregulation of BT's Pricing and the New Fair Trading Requirements' (1997) 8 *Utilities Law Review* 120.

31 There are also tentative proposals for extending collective licence modification procedures in the Review: Green Paper, n 1 above, paras 7.70–72; *Response to Consultation*, n 1 above, para 95, and Department of Trade and Industry, *Licence Modification Procedure: Proposed Changes to the Telecommunications Act 1984* (London: Department of Trade and Industry, 1998).

32 C. Veljanovski, *The Future of Industry Regulation in the UK* (London: European Policy Forum, 1993) 59–60.

 © The Modern Law Review Limited 1999

contract cannot be made for as long a period as five years'.[33] It is easy to defend regulators against the critique implied by the quotation from Veljanovski; for example, governmental claims could also form the basis for 'legitimate expectations' by consumers and employees who were both promised a better deal after privatisation, and legitimate expectations does not work in law to create substantive privileges of this kind.[34] Rather than a regulatory contract having been broken by the regulators, the examples I have given suggest that there is a tension between the actual behaviour of the regulators and the underlying legal model on which their powers are based.

Capture theory and public choice

The model of a bilateral regulatory relationship has, then, been used both as a model for the legal regime adopted and as a source of criticism of regulatory decision-making. It can also be seen in one version of the highly influential theory of regulatory capture which has underlain much analysis of regulation within economics and elsewhere.[35] This proposes that the behaviour of regulators can be predicted from an analysis of the *a priori* interests of regulators and firms and that the result is inevitably that regulation will be established to protect the industry or that the regulator will be captured by the regulated industry. Regulators operate in political markets in which the regulated firm will have most to lose from the regulator's actions therefore and will have the greatest incentive to take action to influence it. As a result it will be able to bid up in the political marketplace by, for example, offering future jobs to regulatory staff, limiting the information available to the regulator or, in extreme cases, by bribery. An influential version is that of the life cycle theory which suggests that a regulator starts off vigorously but then public interest is diverted to other subjects leaving the regulator open to capture.[36] There is a vast body of literature on capture theory and important differences within it; for example, it originated in the Chicago school which may see capture as economically efficient; it cannot be otherwise if it happens as the parties are all assumed to be rational efficiency maximisers and so it affects only distribution of the benefits.[37] Other commentators such as Posner have seen capture as reducing economic efficiency through maintaining barriers to entry and as being costly in itself.[38] The most important point to be made here is however that this early version of capture theory, and indeed most uses of the approach in popular discussion of regulation, concentrates on the bilateral relationship between the regulator and the dominant regulated firm, or at most a small and well organised group of very similar firms. Thus as Stigler put it in the seminal work cited above, '[a] central

33 I. Byatt, *Speech at European Policy Forum on Tuesday 7 May 1996* (Birmingham: OFWAT, 1996) 4–5.

34 See eg *R* v *Secretary of State for the Home Department, ex parte Hargreaves* [1997] 1 WLR 906, and C. Forsyth, '*Wednesbury* Protection of Substantive Legitimate Expectations' [1997] *Public Law* 375.

35 The literature of capture theory is immense and I can only hint at some of its themes here. For useful introductory surveys see Foster, n 6 above, ch 11 and A. Ogus, *Regulation: Legal Form and Economic Theory* (Oxford: Clarendon Press, 1994) ch 6.

36 The seminal statement of capture theory is G. Stigler, 'The Theory of Economic Regulation' (1971) 2(1) *Bell Journal of Economics and Management Science* 1, concentrating mainly on the establishment of regulation to benefit regulated firms. For an early statement of life-cycle theory see M. Bernstein, *Regulating Business by Independent Commission* (Princeton: Princeton University Press, 1955).

37 See Foster, n 6 above, 372–373

38 See *ibid* and eg R. Posner, 'Theories of Economic Regulation' (1974) 5 *Bell Journal of Economics and Management Science* 335–358.

© The Modern Law Review Limited 1999

thesis of this paper is that, as a rule, regulation is acquired by the industry and is designed and operated primarily for its benefit'.[39] If the industry sector consists of more than a single firm the former is assumed to have a single, organisable set of interests which mean that it can be treated as if it were a single firm. This version of capture theory 'singles out a particular interest group – the regulated firms – as prevailing in the struggle to influence legislation, and it predicts a regular sequence, in which the original purposes of a regulatory programme are later thwarted through the efforts of the interest group'.[40]

The approach I have adopted, by contrast, sees the regulator as in the middle of a web of relations, none of which can be assumed to have particular priority and which may well be in fundamental conflict.[41] Therefore capture cannot be predicted as a matter of *a priori* principle. Where regulation is of a multi-enterprise sector it is unlikely that it will be possible to speak of a single industry interest, let alone adequately consider the roles of the workforce or consumers, who may themselves have conflicting interests, for example between domestic and business consumers. This broadening can have both optimistic and pessimistic implications. The optimistic assumption is that of traditional pluralism in that if there are a number of interests involved, and they all have access to regulatory procedures, the interests may balance out and so the more general public interest in the end prevails.[42] The pessimistic implication is that capture may take place by other interests apart from the firm or the industry. On the evidence I gave earlier it is certainly possible to suggest that the early stages of the regulatory process were subject to governmental capture rather than capture by the firm, not in the sense that the regulators were pressurised by ministers but that major decisions, in particular the all-important initial price formulae, were determined directly by government to achieve the political objective of successful privatisation. This is not to say, of course, that industry capture cannot happen; the relationship between producers and the Ministry of Agriculture illustrates powerfully that it can.[43] Instead, I am suggesting that to assume that industry capture is the only, or *ex hypothesis* the most likely form of regulatory evolution, is a drastic oversimplification.

Another approach is apparent in the work of a number of writers from within political science rather than economics. One of the best-known examples is the concept of 'regulatory space' developed by Hancher and Moran.[44] They reject capture theory on the grounds that it assumes a false dichotomy between state and society:

> Economic regulation under advanced capitalism is therefore best conceived as an activity occurring in economies where the public and private are characteristically mixed, where the dominant actors are powerful and sophisticated organizations, and where the biggest firms have taken on many of the features of governing institutions. ... In this world the language of regulatory capture is largely devoid of meaning. ... different institutions have come to inhabit a common regulatory space. The critical question for the analyst of the European regulatory scene is not to assume 'capture', but rather to understand the nature of this shared

39 n 36 above, 1.

40 Posner, n 38 above, 341–342.

41 For an elegant statement of a similar theme see J.Q. Wilson, *The Politics of Regulation* (New York: Basic Books, 1980) ch 10; as he puts it, '[a] single-explanation theory of regulatory politics is about as helpful as a single explanation of politics generally, or of disease' (393).

42 For a useful summary of the pluralist literature on this point see P. Dunleavy, *Democracy, Bureaucracy and Public Choice* (Hemel Hempstead: Harvester Wheatsheaf, 1991) 14–27.

43 G. Cannon, *The Politics of Food* (London: Century Hutchinson, 1987).

44 L. Hancher and M. Moran, 'Organizing Regulatory Space' in Hancher and Moran (eds), *Capitalism, Culture and Economic Regulation* (Oxford: Clarendon Press, 1989).

 © The Modern Law Review Limited 1999

> space: the rules of admission, the relations between occupants, and the variations introduced by differences in markets and issue arenas.[45]

This work offers an extremely fruitful agenda for research and is far more realistic than an approach which concentrates on bilateral relations between regulator and firm. However, it does carry a danger of ending up with a sort of a-theoretical pluralism in which predictions are impossible and every interaction is reduced to a particular, unique instance. On this basis, prediction of regulatory behaviour would be impossible. Perhaps even more importantly we could not make normative decisions about how regulation should interact with its environment or how to make it legitimate, except perhaps by attempting to ensure that all legitimate interests are represented in the regulatory process, an approach which, as I shall suggest below, has difficulties of its own. One way of avoiding these problems is to develop capture theory so that it does not necessarily privilege explanations based on interaction between the regulator and the firm. This has been a theme of importance within capture theory.[46] Indeed, it has produced a distinctive 'Virginian' version of capture theory which can be subsumed within the influential school of public choice.[47] In brief what this work does is to treat politics as a form of marketplace in which all actors behave rationally to maximise utility. Voters use their voting power to extract benefits from the democratic process; politicians use their power to maximise their own wealth or other utility through 'log rolling' and so on; bureaucrats maximise the size of their budgets. This approach would appear to have the advantage of preserving the conception of regulation as a pluralistic enterprise with a number of different purposes and open to influence from a range of interests. However, it raises fresh problems which limit its usefulness.[48] Firstly, if the theory is simply saying that wealth maximisation is one element in politics this is neither new nor particularly helpful; it is familiar from muckraking journalism let alone more elevated sources, and does not help in analysing *where and in which circumstances* wealth maximisation can be used successfully to predict regulatory behaviour. If it is saying that wealth maximisation is the *only* valid explanation of political behaviour (and to have predictive power it presumably must be saying this) it is palpably false; prestige, an easy life and even altruism and public service are well documented sociologically as parts of political life.[49] A further possible response is to expand the definition of utility to include non-monetary utility such as prestige, 'an easy life' and 'feeling good' rather than maximisation of material wealth, but if everything can count as utility all predictive value is taken away from the theory and we are back with the sort of pluralist contingency this theory was designed to overcome. As Foster puts it:

> though this theory involves a gain in realism it loses in predictive power. The proponents of such a theory easily find themselves in a situation where, whatever the regulatory outcome, it can always be rationalized as the achievement of a political equilibrium between the various interests ... Because one does not know what weight to give in each case to money,

45 *ibid* 276. For an application of these ideas see Scott, Hall and Hood, n 17 above, and A. Melville, 'Power, Strategy and Games: Economic Regulation of a Privatized Utility' (1994) 72 *Public Administration* 385.

46 See in particular S. Peltzmann, 'Towards a More General Theory of Regulation' (1976) 19 *Journal of Law and Economics* 211 and Posner, n 38 above.

47 For a summary of the 'Virginia' approach see Foster, n 6 above, 384–388; for an excellent critique of the enormous public choice literature see Dunleavy, n 42 above.

48 cf Wilson, n 41 above, 361–363.

49 For a sustained attack on public choice theory on this and other grounds, see M. Kelman, 'On Democracy-Bashing: A Skeptical Look at the Theoretical and "Empirical" Practice of the Public Choice Movement' (1988) 74 *Virginia Law Review* 199.

© The Modern Law Review Limited 1999

votes, ideology and other factors, some of which cannot be measured at all, one has no measures of comparative efficiency.[50]

A more useful revisionist approach to capture theory has concentrated on the complexity of the institutional and political context in which regulation may take place, and in this sense has much in common with the regulatory space approach.[51] Thus for example Levine and Forrence have stressed the limitations on capture which result from issues being moved onto the public agenda, whilst Makkai and Braithwaite emphasise multidimensional pressures towards different types of capture which can be countered by, for example, granting third party participation rights. This approach is weak in asserting a base for substantive normative principles of regulation.[52] However, it is valuable in providing the basis for a procedural approach to regulation. This leads us to a different approach which I shall now discuss before asking whether it is possible to move beyond proceduralism towards substantive principle.

Stakeholder theory

A quite different approach to understanding regulation is now becoming fashionable in the form of stakeholder theory. This would appear to offer a way out of the problems I have just raised both predictively through identifying the key interests which may influence regulation and normatively through suggesting the interests which should participate in the regulatory process. It would seem to present a possible solution to the criticisms of capture theory that I have made above as, rather than seeing regulation as a bilateral relationship between regulator and firm, regulation encompasses a network of relations involving the dominant firm, its competitors, its consumers and others such as employees and suppliers. Indeed, a commentator on utility regulation under the aegis of the influential Institute for Public Policy Research has explicitly advocated 'a stakeholder approach to regulation' in which government sets a more explicit framework for the regulators, one which will reflect both the national interest and those of different stakeholder groups.[53] Thus regulators themselves 'should pay explicit attention to the interests of different stakeholder groups and adjust their regulatory instruments to ensure that outcomes are not inconsistent with a desirable balance between them'.[54] Government should also play a part through setting broad sectoral policies in mission statements agreed with regulators. The relevant stakeholders are to be business and residential customers, shareholders, utility managers, market entrants, suppliers and other companies dependent on utility industries, and employees in the affected sectors.[55] Significantly, there has been some regulatory support for this view.[56] Moreover, the Review also shows the influence of this approach; 'if regulation is to win long-term acceptance, and thus provide a stable framework for

50 n 6 above, 387.

51 See notably M. Levine and J. Forrence, 'Regulatory Capture, Public Interest and the Public Agenda: Toward a Synthesis' (1990) 6 *Journal of Law, Economics and Organization* 167, and T. Makkai and J. Braithwaite, 'In and Out of the Revolving Door: Making Sense of Regulatory Capture' (1995) 12 *Journal of Public Policy* 61.

52 See eg Levine and Forrence *ibid* 181–182 where the substantive analysis of policies is left to 'a broader philosophical debate' or to notional ratification by an informed polity.

53 D. Souter, 'A Stakeholder Approach to Regulation' in D. Corry, D. Souter and M. Waterson, *Regulating Our Utilities* (London: Institute for Public Policy Research, 1995).

54 *ibid* 45.

55 *ibid* 35–53.

56 I. Byatt, n 33 above, 6.

© The Modern Law Review Limited 1999

business, it must be fair, and seen to be fair, by all those with a stake in the utilities'.[57] The influence of stakeholder theory also appears in the frequent references to maintaining a fair balance between consumer and shareholder, the role of social and environmental guidance and increasing transparency.[58]

A stakeholder approach also seems to provide the basis for an alternative conception to the bilateral regulatory contract approach which I have seen as at the heart of the problems. Thus there have been moves in areas of corporate governance and company law which have conceived of corporate governance as analogous to politics, both in the sense that it offers indirect forms of interest representation within the firm, and in the sense that relations are not seen as those based on bilateral contracts but on relational contracts establishing long-term relationships between a number of parties, subject to their own procedures for change, or on a nexus of contracts between them.[59] This does not necessarily imply a stakeholder approach as the concern may only be with the relations between the company and its members and between the members themselves; however it does show that even a contractual model may be far more complex than the bilateral one criticised above. A further step is to make an analogy with public law based on the social importance and power of the modern company; to quote Stokes, 'the large company tends to collapse the distinction between private and public power. The company can thus plausibly be viewed as a miniature state. All those affected by its decisions are citizens of that state and should therefore be involved in making those decisions'.[60] As a result a stakeholder approach is necessary to open up accountability to the wide range of interests affected.[61]

If this opening up of the bilateral relationship can be valid when assessing internal controls in the firm, surely it can be even more so in understanding the basis of external controls in the form of regulation? Indeed, there would seem to be two reasons for suggesting that it would be even more valid. Firstly, the essential nature of the services provided by utilities means that their decisions affect particularly important social interests; I shall deal with the question of how we identify those interests at the end of this article. Secondly, one forceful critique of the stakeholder approach, made both by right and left, is that it underplays the role of the market environment in which the enterprise operates. Although enterprises may engage in window dressing about responsiveness to stakeholders, in the end the market imperative to maximise profit will always prevail. Thus '[i]n reality, the dynamics of capitalist society are such that it is ultimately presided over not by people but by the market: neither the capitalist entrepreneur nor the paid manager *controls* their own enterprises, let alone the economy as a whole, except in a limited sense'.[62] This is questionable in itself.[63] Even if it were to be accepted,

57 Green Paper, n 1 above, para 1.5.

58 See eg Green Paper, n 1 above, paras. 2.9, 5.1–2, 7.18; *Response to Consultation*, n 1 above, paras 6, 96.

59 R. Drury, 'The Relative Nature of a Shareholder's Right to Enforce the Company Contract' (1986) 45 *Cambridge Law Journal* 219. See also the seminal work in this area, M. Stokes, 'Company Law and Legal Theory' in W. Twining (ed), *Common Law and Legal Theory* (Oxford: Blackwell, 1986) 155.

60 Stokes, *ibid* 178.

61 See (from an enormous literature) J. Kay and A. Silberston, 'Corporate Governance' (1995) 153 *National Institute Economic Review* 84 and for some reservations, P. Ireland, 'Corporate Governance, Stakeholding, and the Company: Towards a Less Degenerate Capitalism' (1996) 23 *Journal of Law and Society* 287, with a reply by D. Campbell, 'Towards a Less Irrelevant Socialism: Stakeholding as a "Reform" of the Capitalist Economy' (1997) 24 *Journal of Law and Society* 65. For an attempt to apply a stakeholder approach see Royal Society of Arts, *Tomorrow's Company: The Role of Business in a Changing World* (London: Royal Society of Arts, 1995).

62 Ireland, *ibid* 305 (emphasis retained). For a critique from the right on similar grounds see eg M. Friedmann, *Capitalism and Freedom* (Chicago: University of Chicago Press, 1962).

63 See Campbell, n 61 above.

© The Modern Law Review Limited 1999

however, as I outlined at the beginning of this article, the operation of the utility regulators cannot be seen as constrained to mimic markets; other rationales such as some types of social regulation inevitably play a part in their decisions. This has been so even in areas such as the development of competition in domestic telecommunications and energy, which are at the heart of their economic tasks. Moreover, the statutory duties which apply to the regulators cannot themselves be seen as implementing any single logic of maximising profit for regulated companies, nor will the changes proposed in the Review do so.[64] Thus, even if an individual firm is constrained by a fundamental market logic (and this is by no means as simple as it may sound), a regulator need not be.

An apparent strength in the stakeholder approach is that it is not merely descriptive but has a strong normative dimension; not only do stakeholders have interests, but there is a moral obligation on companies or regulators to take those interests into account. This could provide a framework for assessing the legitimacy of regulatory decisions and so fill the gap in this legitimacy identified above. There may be individual reforms such as opening appeal rights to interests other than the regulated firm which we can make uncontroversially through applying stakeholder theory (though this is not proposed in the Review).[65] However, there is a serious problem in utilising stakeholder theory as a general theory for the prediction of regulatory behaviour and for making normative decisions of regulatory design. This was noted by the Director General of Water Services after cautious commendation of the stakeholder approach; '[t]he stakeholder approach does not tell us what the right balancing of interests is'.[66] Thus, whilst advocating opening up regulation to a wider range of interests, stakeholder theory provides no criteria for determining which interests are to count and how much they are to count. This depends on some higher, unarticulated theory or set of political values. A controversial example would be that of the workforce. So far, regulatory decision making has played minimal attention to the interests of the workforce; indeed, the interest of employees is the one interest conspicuously missing from those reflected in the statutory duties of regulators, and again there is no suggestion for change in the Review.[67] This need not always be so: earlier regulation gives us the example of the 'basic price system' of price control used for the gas industry from 1920 in which extra dividends payable in certain circumstances had to be matched by equivalent payments to employees through co-partnership schemes or bonuses.[68] We could argue that the workforce should have a role in the regulatory process for the public utilities, but why? Souter in the work cited above bases the involvement of employees as stakeholders on the fact that 'they have much to lose if things go wrong' and that 'competition and price regulation have also put considerable pressure on utility managers to cut costs, particularly labour costs'.[69] This appears compelling, but it could also be applied to, for example, the employees of supplier and competitor firms; where do we draw the line between stakeholders and non-stakeholders? More seriously, this approach gives us no indication of the relative weight to be attached to different stakeholder interests; it is hardly inconceivable that employee and consumer interests will be in direct

64 Prosser, n 4 above, ch 1; cf J. Ernst, *Whose Utility? The Social Impact of Public Utility Privatization and Regulation in Britain* (Buckingham: Open University Press, 1994) 60.

65 For an example see Prosser, n 4 above, 301–304.

66 Byatt, n 33 above, 6.

67 cf the duty of company directors to consider the interests of employees in Companies Act 1985, s 309.

68 *The Gas Industry: Report of a Committee of Enquiry*, Cmd 6699 (1945) paras 45–46.

69 n 53 above, 43.

© The Modern Law Review Limited 1999

conflict and simply to point to both as stakeholders hardly takes us any further in resolving that conflict. Finally, there is something to be said for the argument that proliferating the number of stakeholders will result in a confusing melange of conflicting objectives with none adequate to offer realistic guidance; as has been suggested in the context of corporate governance

> 'Making bosses accountable to many stakeholders might make them accountable to none, as there would be no clear yardstick for judging their performance' ... This problem is familiar from Britain's experience with nationalised industries, where the multiplicity of goals led to confusion between the concerns of government and those of industry.[70]

The easiest way out of the dilemma of who is to count as a stakeholder would be to recognise as stakeholders those whose interests are affected. This merely moves the difficulties one step back, however, for we need to determine which interests are worth protecting and to develop some order of ranking between them. Even in corporate governance there is still considerable controversy over which interests count as stakeholder interests, even to the extent of disagreement on whether shareholders themselves have a central role.[71] Moreover, there is also no consensus on the basic objectives which stakeholding is to promote; are they concerned essentially with assisting profit maximisation or partially replacing it with broader social concerns?[72] Such difficulties are likely to be even greater in regulation where there is no central organising nexus for different interests like that of the company. Thus stakeholder theory on its own does not take us beyond regulatory pluralism; it needs supplementing by other theory which might help in identifying and ranking interests.

Proceduralism in regulation

The conclusion that I have just drawn may seem to be too harsh. One advantage of stakeholder theory may be alert us to the need for participative procedures in regulation. Thus even if we cannot precisely identify stakeholders or rank the interests they represent, at least we can be open to the need to ensure that the regulator has access to information from as wide a range of interests as possible in reaching decisions. This is in contrast to capture theory which has generally been pessimistic about the role of participative procedures; they are seen as inevitably leading to domination by the regulated industry as the only participant with the resources to ensure that its own view prevails through them at the expense of efficiency-maximisation or of the public interest. For example, in the pioneering work cited above, Stigler suggests that

> the procedural delays required of public processes are costly. The delays which are dictated by both law and bureaucratic thoughts of self-survival can be large ... Finally, the political process automatically admits powerful outsiders to the industry's councils. It is well known that the allocation of television channels amongst communities does not maximise industry revenue but reflects pressures to serve many smaller communities. The abandonment of an unprofitable rail line is an even more notorious area of outsider participation.[73]

---

70 Kay and Silberston, n 61 above, 93. The quotation is from M. Bishop, 'A Survey of Corporate Governance' (1994) 330 *The Economist* 53. For detailed discussion of the nationalised industries problem see Foster, n 6 above, ch 3.

71 Kay and Silberston, n 61 above.

72 See eg Ireland, n 61 above, esp 300–306, 310–315

73 Stigler, n 36 above, 6. The work of Levine and Forrence and of Makkai and Braithwaite (n 51 above) is an important exception to this hostility to participative regulation.

© The Modern Law Review Limited 1999

Rather than seeing the involvement of community interests as 'notorious', stakeholder theory would see it as highly desirable.[74] This emphasis on the virtues of participative procedures also has the advantage of being close to the actual practice of utility regulators in the UK, and of course in the United States. In the former, the procedures adopted for the making of key decisions have generally been much more open to a range of interests than the law requires, whereas in the latter there is an enormous body of administrative law designed precisely to permit such interest representation.[75]

The limits of substantive legal regulation have become the subject of a large body of literature recently. Of particular importance has been that drawing on the theory of autopoiesis, theory which is highly complex and which cannot be rehearsed in full in this article but which has regulatory implications which should be examined here.[76] One starting point of the theory is closely related to many current critiques of the utility regulators; as King has put it,

> [t]here have proved to be a number of problems ... with the Utopian vision of law as an effective regulator of social behaviour. Not least of these was the growing evidence of law's failure to fulfill the high hopes that had been expected of it. Attempts to regulate social systems through law often seemed to work effectively for a short time and then would produce consequences which were unanticipated by legislators and would-be reformers.[77]

This is largely attributable to the fragmentation of modern societies into differentiated functional systems and sub-systems, including law, politics and the economy. These sub-systems are cognitively open but normatively closed; this means that, to quote King once more, '[t]he normative communications of other systems cannot simply be reproduced by law as legal communication. They first have to be reconstructed as law if they are to become accepted as law, and this reconstruction process may well give rise to unforeseen distortions and reductions to the meaning of the original communications as they were formulated in the political or economic systems'.[78] Similar communication problems also occur in the attempted use of law to influence the behaviour of economic systems. The problem is illustrated in Teubner's discussion of price control. If law imposes a price freeze:

> [t]his is normally understood as a clear case of the law intervening directly in the economy. From the point of view of autopoiesis, however, it is merely an act of observation ... Through the device of the price-control norm, the law is merely observing its own operations, imagining all the while that the economy functions in such and such a way ... From the moment it is promulgated to the moment it is implemented, our price control

74 See I. Ayres and J. Braithwaite, *Responsive Regulation: Transcending the Deregulation Debate* (Oxford: Oxford University Press, 1992) ch 3.

75 For the UK see Prosser, n 4 above, 83–86, 113–115, 144–147, 177–178, 198–199, 259–266, 296–297; for analysis and critique of the US 'interest representation' model see R. Stewart, 'The Reformation of American Administrative Law' (1975) 88 *Harvard Law Review* 1669–1813

76 For a very clear introduction to autopoiesis see M. King, 'The Truth about Autopoiesis' (1993) 20 *Journal of Law and Society* 218; for an earlier example with some direct relevance to regulation see G. Teubner, 'Substantive and Reflexive Elements in Modern Law' (1983) 17 *Law and Society Review* 239, and for later development eg G. Teubner, 'Juridification: Concepts, Aspects, Limits, Solutions' in G. Teubner (ed), *Juridification of the Social Spheres* (Berlin: de Gruyter, 1987) and his *Law as an Autopoietic System* (Oxford: Blackwell, 1993). For a direct application to regulatory concerns see J. Black, 'Constitutionalising Self-Regulation' (1996) 59 MLR 24. The literature is once again enormous; for an interesting criticism see J. Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (Cambridge: Polity Press 1996) 52–56.

77 King, *ibid* 222 (footnote omitted).

78 *ibid* 227; for a summary see also Teubner, n 76 above (1993) 69–71.

 © The Modern Law Review Limited 1999

> consists solely in cognitive and normative operations which are played out within the law and which cannot be transferred to the economy ... this procedure does not lead from legal conceptions of the economy to the reality of the economic system itself.[79]

This would appear very apposite to analysis of the problems of regulators attempting to control excess profits through price control, or to ensure that the water companies and Railtrack properly use profits for investment rather than passing it on in dividends.

Autopoiesis, then, posits the existence of severe restrictions on the direct regulatory capacities of law. It is worth noting, incidentally, in the context of this article, that it also posits serious restrictions on the ability of law to implement the logic of the economic system; the economics of law model 'leads to a systematic failure to note the capacity of the legal system to select economic input'.[80] As regards law's regulatory limitations, growing juridification can lead to Teubner's 'regulatory trilemma' in which law is either ignored by other systems, or destroys those systems' traditional and appropriate norms of behaviour, or is itself disintegrated by the pressures imposed on it by other systems.[81]

Is the implication of this analysis an end to law as a regulatory device, an end to intervention by public authorities in economic systems, indeed an end to regulation? This is not the solution proposed. Rather than abandoning law, the proposal is for 'reflexive law'. It has some characteristics in common with the stakeholder approach discussed above; in an early statement of the theme it was described in the context of contract law as follows:

> Reflexive law ... seeks to structure bargaining relations so as to equalize bargaining power, and it attempts to subject contracting parties to mechanisms of 'public responsibility' that are designed to ensure that bargaining processes will take account of various externalities. However, within the limits of the arena that has been so structured, the parties are free to strike whatever bargains they will. Reflexive law affects the quality of outcomes without determining the agreements that will be reached. Unlike formal law, it does not take prior distributions as given. Unlike substantive law, it does not hold that certain contractual outcomes are desirable.[82]

Reflexive law thus represents a new proceduralism; rather than determining outcomes directly, it fosters the development of structures for reflexion within other social subsystems, including through compensation for inequalities in power and information. Thus 'instead of directly regulating social behaviour, law confines itself to the regulation of organization, procedures and the redistribution of competences'.[83]

Once more we see strong echoes of some of the issues concerning UK utilities regulation here; most notably in the greater proceduralisation of regulation through the incorporation of interest groups and in growing interest in 'self regulation' or 'enforced self regulation' as a regulatory strategy.[84] Two reservations have however to be entered in relation to the use of this work as the sole approach to analysis of this regulation. The first is that it is doubtful to what extent the process of juridification has advanced in this area of UK administration. It is certainly true that judicial review and other forms of legal challenge have become more

79 Teubner, *ibid* 77–79.
80 *ibid* 57.
81 Teubner, n 76 above (1987) esp 19–27; Black, n 76 above, 47.
82 Teubner, n 76 above (1983) 256 (footnote omitted).
83 Teubner, n 76 above (1987) 34; see also Teubner, n 76 above, (1983) 270, 275, 277 and Black, n 76 above, 44, 46.
84 See Black, n 76 above, and Ayres and Braithwaite, n 74 above, ch 4.

© The Modern Law Review Limited 1999

common.[85] Nevertheless, in some areas of regulation at least the tests applied by the courts have been notably open textured, leaving a considerable power to regulators to set priorities and balance competing interests.[86] Moreover, the legislation also permits considerable regulatory discretion, even to the extent of largely leaving the ranking of regulators' basic objectives to the regulators themselves; the re-ordering of statutory duties proposed in the Review will not do much to change this.[87] Thus the excesses of detailed substantive prescription by law have been largely avoided in the utility regulation sector, especially compared to some areas of local government where juridification has been used as a political tool in battles between central government and local authorities.[88] Moreover, juridification means more than an increase in the number of legal rules. It implies that regulatory law is used in which, rather than general rules, particularistic requirements are made of conduct to achieve substantive social ends; in Weberian terms, law is materialised.[89] Yet, as Scott has pointed out, at least in the field of telecommunications current moves seem to be towards dematerialisation through the replacement of particularist norms by general norms, reflecting both the development of multi-enterprise markets and the role of European requirements.[90] It would thus, so far at least, be an exaggeration to talk of the juridification of utilities regulation showing that substantive regulation through law is unworkable. Examples of difficulties of regulatory implementation in such areas as price control have certainly existed; however, there may be more contingent explanations for this, in particular the setting of the first generation of price controls over-generously for the companies by Government in order to assist the privatisation process, and the expectation of over-long periods between periodic price reviews. It is certainly possible to point to areas in which substantive regulation has been at least partially successful, for example in the setting of service standards and the decline of disconnection rates. Even the later generations of price controls are now more demanding and are certainly strongly resisted by regulated companies. In this respect there is a difference between setting a price control for a particular enterprise and Teubner's example of a price freeze for a whole economy, where information and enforcement problems are indeed likely to make substantive legal prescription of only minimal effectiveness.

The second problem in applying the approach of reflexive law as the sole basis for our understanding of utilities regulation is that it leaves unanswered the problem underlying much of this paper; that of the substantive values which should underpin utility regulation. Of course, for reasons mentioned above, a reflexive law approach would reject the proposition that an economic approach to regulation has or should have priority in the political and legal sub-systems of regulation. But what would fill the void thus created? As Black has asked, '[i]s reflexive law simply a technique of intervention, a vessel into which any substantive purpose can be poured, or does it restrict not only the means by which substantive goals can be

85 C. Scott, 'The Juridification of Regulatory Relations in the UK Utilities Sectors', in J. Black, P. Muchlinski and P. Walker (eds), *Commercial Regulation and Judicial Review* (Oxford: Hart Publishing, 1998). A key decision was *Mercury Communications Limited* v *Director General of Telecommunications* [1996] 1 WLR 48

86 This has been particularly the case in the regulation of broadcasting: *R* v *Independent Television Commission, ex p TSW Broadcasting Ltd* [1996] EMLR 291 (HL); *R* v *Independent Television Commission, ex p Virgin Television Ltd* [1996] EMLR 318.

87 See Prosser, n 4 above, ch 1.

88 See M. Loughlin, *Legality and Locality – The Role of Law in Central-Local Government Relations* (Oxford: Clarendon Press, 1996) ch 7.

89 Teubner, n 76 above (1987) 10–19.

90 Scott, n 22 above.

© The Modern Law Review Limited 1999

pursued, but those that can be pursued at all?'[91] It seems clear that a role for substantive values is envisaged in reflexive law; to quote Teubner,

> [t]he term 'proceduralization' of law is easily misunderstood from a technical-legal point of view. To interpret it as a recommendation to a 'post-interventionist' legislator to do without substantive legal norms and rely exclusively on procedural law instead would be to miss the point ... substantive legal norms remain indispensable. It is only that the process of their production and justification has to give way to a 'socially adequate' proceduralization.[92]

Yet this continuing role for substantive values creates further problems similar to those identified in the discussion of stakeholder theory above; if we want to use participation of interests to determine the content of substantive standards, this requires a means of identifying which interests count, and how we are to balance them against each other. Such considerations have led to suggestions of the possibility of using reflexive law to assure constitutional rights; '[h]owever, whether we can simply download any political or constitutional theory or theory of rights we want is uncertain; moreover, the nonindividualistic basis of the theory of reflexive law means that insofar as they are focused on individual rights, current constitutional theories or aspects of them would have to be reformulated'.[93]

Conclusions: beyond proceduralism?

I have argued that both the legal model adopted for regulation of public utilities in the UK and the leading critical school of capture theory are inadequate for a proper understanding of regulation, and for prescription about how to improve regulation in practice. In both cases this is due to an unjustified privileging of the relations between regulator and enterprise as the subject of study whilst neglecting the pluralism actually characteristic of regulation. In terms of description and positive analysis of regulation, the regulatory space approach would appear to have much more to offer simply because it does not privilege any particular relationship in explaining or predicting how regulation will operate. However more is needed to provide analytical and evaluative purchase and normative principles as a basis for regulatory reform, as a guide for regulators and indeed for government in exercising its regulatory responsibilities. The answer might seem to be stakeholder theory, but this suffers from vagueness on the nature and weight of interests to be recognised as stakeholders. Proceduralism similarly would seem to have advantages in encouraging a more sensitive regulatory style but also leaves unanswered key questions of determining legitimate interests and their role in participation. Is there any way we can go beyond these theories to offer some solutions to these central question of regulation?

It has been suggested above that we abandon any pretence that a distinctive body of regulatory theory can be developed at this level; instead the better approach is simply to treat regulation as a sub-branch of government and so to return to political and constitutional theory. This will prove useful in two respects. The first is in setting the substantive principles which should underlie regulation; these may include those based on economic efficiency and those based on more distributive goals. The second is to identify the interests which should be taken into account in

91 n 76 above 48–49.

92 Teubner, n 76 above (1993) 66–67.

93 Black, n 76 above, 50 (footnote omitted). Her discussion of this issue at 48–55 is a particularly illuminating one, cf C.B Graber and G. Teubner, 'Art and Money: Constitutional Rights in the Private Sphere?' (1998) 18 OJLS 61.

© The Modern Law Review Limited 1999

the procedural process and to give some criteria to be used in balancing them. Of course, both can be accomplished by examining the positive legal duties which apply to regulators, and the Review proposals may make this easier through tidying these up to some degree. Yet the statutes are not an exhaustive source of regulatory principle; if they were there would be no need for debate about how regulators should decide. In addition, the regulatory space should be seen as operating within a set of boundary constitutional principles. In other nations this would not be controversial; one has only to think of the importance of the US due process clause in shaping regulation there, or the development of the concept of public service in French constitutional and administrative law; other examples are familiar from the protection of public service broadcasting as a constitutional right in a number of jurisdictions.[94] In the UK, given the lack of a relevant set of determinate constitutional principles, a more speculative approach is necessary. One way of undertaking this would be an explicitly rights-based approach in which regulation is seen as operating in a framework of rights, including social and economic rights.[95] It is important to note that a rights based approach of this kind will not necessarily contradict the adoption of economic goals including market liberalisation. As Dworkin has argued, there may be a strong argument in many circumstances for markets as means of implementing rights on the ground that markets provide the best means of treating people as equals by ensuring that they take responsibility for the true costs of their actions; '[s]omeone who abstains from some act on the ground that it would cost his neighbour more than it would benefit him takes his neighbour's welfare into account on equal terms with his own; a duty to act in that way might be thought to rest on some egalitarian basis'.[96] Alternatively the rights-based arguments may provide a justification for not implementing economic efficiency-based policies on the ground that we do not enter markets as equals and so such policies will defeat other values, such as ensuring access to essential social facilities on the part of the worst-off members of society.[97]

A more self-contained approach might be to take the participative or proceduralist emphasis as given, and to analyse the conditions which are implied for it to operate. This comes close to the approach taken in recent work by Jurgen Habermas.[98] To summarise the relevant aspects of some extremely rich (and dense) theorising, he accepts (for reasons of efficiency as well as of rights) that what is required is 'a "democratization" of the administration that, going beyond special obligations to provide information, would supplement parliamentary and judicial controls from within'.[99] This would appear to align him with proceduralism, but he also develops a scheme of rights which are required to institutionalise democratic processes of discourse. These include the traditional negative liberties, due process rights and rights of political participation but also social welfare rights which are necessary for any effective participation; '[b]asic rights to the provision of living

94 See R. Craufurd Smith, *Broadcasting Law and Fundamental Rights* (Oxford: Clarendon, 1997).

95 See notably A. Gewirth, *Reason and Morality* (Chicago: University of Chicago Press, 1978); *The Community of Rights* (Chicago: University of Chicago Press, 1996), and for applications see D. Beyleveld, 'The Concept of a Human Right and Incorporation of the European Convention of Human Rights' [1995] *Public Law* 577 and D Beyleveld, and C. Villiers, 'A General Right to a Minimum Wage in English Law: An Argument from Generic Consistency' (1997) 17 *Legal Studies* 234.

96 R. Dworkin, *Law's Empire* (London: Fontana, 1986) 295 and see generally ch 8.

97 See eg A. Coote, *The Welfare of Citizens. Developing New Social Rights* (London: Institute of Public Policy Research, 1992).

98 Habermas, n 76 above. For an attempt to apply his earlier work in a related context see T. Prosser, 'Towards a Critical Public Law' (1982) 9 *Journal of Law and Society* 1.

99 n 76 above, 440.

© The Modern Law Review Limited 1999

conditions that are socially, technologically, and ecologically safeguarded, insofar as the current circumstances make this necessary if citizens are to have equal opportunities to utilize the civil rights ...'.[100] This concept of rights as presupposed by participation can operate in two ways: on a micro-level as presupposed by democratic and participative procedures adopted by government and regulators, but more importantly on a macro-level of the rights presupposed by participation in the broader society in which regulation takes place. Such rights could include access to basic services such as those necessary for communication, access to a common culture through public service broadcasting and to basic means of subsistence in the form of energy and water. Moreover, a ranking of rights in this way might provide us with a means of ranking the interests of different stakeholders. Thus where rights are best met by market provision this might be where shareholder interests and those of consumers as market actors are all that need to be recognised; where social and economic rights come into play a broader conception of stakeholder would apply to include those who are disadvantaged in the market place but require access to public services and those who represent them. Less convincingly, it could be argued that the lack of representation of employees in the regulatory process is justified because an open labour market together with protection against such disadvantages as unfair dismissal are the means by which their rights and interests are protected, and so they need not be recognised as stakeholders in the regulatory context, though of course they may be in others. For example, it could be argued that protection is afforded against undue pressure for changes in the workforce and its conditions of employment through the Transfer of Undertakings Regulations.[101] However, if these protections are inadequate and if regulatory decisions create specific pressures on the workforce this may justify recognition of employees' interests in regulatory decisions.

A second approach which may simplify matters somewhat is to seek regulatory principle not only in political philosophy but as part of the international obligations to which government is party.[102] This is not entirely separate from the first approach but does give an opportunity to find relatively concrete norms as a basis for policy decisions. Thus it could be argued, for example, that the Council of Europe Social Charter could form the basis for developing social rights which could include the right to participate in essential services.[103] Much more important have been European Community obligations. In the utilities sector these have been concerned both with liberalisation and with protecting certain social rights within the process; this is well advanced in telecommunications but over the next few years is likely fundamentally to affect the energy sector also.[104] The role of European obligations is also becoming fundamental in the regulation of broadcasting and in particular of the new media.[105]

Two objections may be made to the approach suggested here. The first is a matter of principle; decisions of this kind are best left to elected governments on

100 *ibid* 123.

101 The Transfer of Undertakings (Protection of Employment) Regulations 1981, SI 1981 no 1794.

102 See Beyleveld, n 95 above and Beyleveld and Villiers, n 95 above, for examples.

103 N Lewis and M. Seneviratne, 'A Social Charter for Britain' in Coote (ed), n 97 above. See also N. Prouvez, 'The European Social Charter- an Instrument for the Protection of Human Rights in the 21st Century?' (1994) 53 *The Review* (International Commission of Jurists) 30.

104 For a summary see J. Pelkmans, 'Utilities Policy and the European Union' in Centre for the Study of Regulated Industries, *Regulatory Review 1997* (London: Centre for the Study of Regulated Industries, 1998) 111.

105 S. Verhulst and D. Goldberg, 'The European Institutions', in Goldberg, Prosser and Verhulst, n 11 above.

© The Modern Law Review Limited 1999

the basis of the argument from democracy mentioned early in this article and so are inappropriate for regulators. There is some force in this, but for a rather different reason from that of democratic legitimacy; as was argued earlier, governments are not necessarily more accountable than are regulators. However, governments are in a better position to take a view on the relevant matrix of rights and interests as a whole than is a regulator with specialist expertise in a particular sector. This argues for a form of policy guidance from government to regulators of the sort proposed in the Review, although this will not remove all regulatory discretion in particular instances. The second objection is a matter of practicality; it is difficult enough for regulators to be 'governments in miniature' without asking them to become philosophers writ large! The first response to this objection is, as suggested above, that the sort of rights I am proposing are not to be applied by regulators alone. In principle, it is possible to distinguish different levels of normative analysis which could reflect a division of labour; Gewirth thus distinguishes, firstly, the constitutional and legislative level of enacting a constitution and laws which support the fulfillment of social and economic rights, and secondly, how they are best and most efficiently to be implemented.[106] In some cases this has been reflected in the actual distribution of responsibilities; for example, in the United States the important reforms of the Telecommunications Act 1996 relating to universal service set out six principles on which the regulatory authorities were to base their policies, whilst leaving details of implementation and the development of additional principles to the authorities themselves.[107] However, in the UK regulatory duties have been highly open-ended and this has made it inevitable that decisions have had to be taken on matters of principle by the regulators; the reforms proposed in the Review may go some way to improving this.

The second response to the objection based on impracticality is that there are signs that utility regulators are increasingly doing something similar in practice to what has been proposed here. The strongest example is that of universal service in telecommunications. This became a regulatory concern early on, but was then based very much on the application of solely economic principle; at the time of the 1986 price review the then Director General of Telecommunications accepted that some restraint on line rentals could be justified by the argument that the more customers who joined the telephone service the more valuable it would be for all customers but felt that it would not 'be proper for me to attempt to redistribute wealth by my policy on telephone pricing'.[108] During the mid-1990s, however, much more work was carried out by the regulator on universal service and this encompassed more clearly social concerns; the first consultative paper stated 'OFTEL has a responsibility to promote the interests of all consumers, including people with a disability or of pensionable age and a general duty to regulate in the public interest. A fair and equitable balance needs to be struck between the interests of different groups'.[109] The final consultative paper, whilst repeating the economic argument, stated:

> [t]he concept of universal service is based on the premise that telecommunications services now play such a fundamental role in our society that all people, whoever or wherever they

106 Gewirth, n 95 above (1996), 327.

107 Telecommunications Act 1996, 47 USC §§ 151 *et seq*, s 254(b), for the outcome of the regulatory process see Federal Communications Commission, *Report and Order in the Matter of Federal-State Joint Board on Universal Service* (Washington: FCC 97–157, 1997).

108 OFTEL, *Review of British Telecom's Tariff Charges* (London: OFTEL, 1986)

109 OFTEL, *A Framework for Effective Competition* (London: OFTEL, 1994) para 12.1; for a more detailed account of the consultation process see Prosser, n 4 above, 78–83.

© The Modern Law Review Limited 1999

> are, must have access to a certain basic level of telecommunications facilities and services if they are to participate fully in modern society. The concept of universal service is well established in the UK, having been a feature of telecommunications service since long before BT was privatised. It provides a safety net – ensuring that people who may otherwise miss out get access to a basic level of service. It is also a dynamic concept and one which will evolve over time.[110]

This philosophy has also been strongly influenced by European Community developments which have themselves developed a sophisticated concept of universal service.[111] Thus we see a progression from an argument based on economic efficiency to one based on a right to participate and a right to a basic service.

In conclusion, my argument has been that study of utility regulation suggests that such regulation is an essentially open process and cannot, and indeed should not, be reduced to any particular logic, economic or otherwise. The bilateral approach and capture theory deny this complexity and so remove the richness from regulatory activity; because of this they are of limited utility either in predicting regulatory behaviour or in providing normative principles of how regulators should act. Public choice and stakeholder theories suffer from similar problems. The simple answer might be to abandon any search for regulatory theory and say that regulation is merely a part of politics and treat it as such. This does not however mean that it is arbitrary or simply a matter of contingent compromises between interests, for some ways do exist for developing more sophisticated regulatory principles without artificially pushing the regulatory process into simplistic theoretical straight jackets.

110 OFTEL, *Universal Telecommunications Services: Proposed Arrangements for Universal Service in the UK from 1997* (London: OFTEL, 1997).

111 See eg Commission of the European Communities, *Developing Universal Service for Telecommunications in a Competitive Environment*, COM(93)159 (final); European Parliament and Council Directive 98/10/EC on the application of open network provision to voice telephony and on universal service for telecommunications in a competitive environment, OJ L 101, 1.4 1998.

© The Modern Law Review Limited 1999

Reconsidering Disgorgement for Wrongs

Sarah Worthington*

Nobody should be permitted to profit by wrongdoing: this sentiment has compelling intuitive appeal. Despite this, profits disgorgement[1] (or stripping the defendant of ill-gotten gains) turns out to be a remedy with surprisingly limited application. This may soon change. Both the Law Commission[2] and the judiciary[3] have indicated support for wider recognition of the remedy; so too have academics.[4] However, few would suggest the remedy ought to be available for all profit-generating wrongs,[5] and here lies the difficulty. As yet no theory satisfactorily explains which wrongs should give rise to disgorgement and which should not. Sometimes the focus has been on the character of the wrong; at other times it has been on the moral culpability of the wrongdoer. More importantly, and more worryingly, no theory explains when the remedy should strip the defendant of every penny of the ill-gotten gain and when something less – generally 'expenses saved' or 'use value' – should suffice. This article re-assesses existing law and suggests that an alternative analysis may provide some answers.

Put briefly, this article makes two claims. The first is based on an examination of existing case law. It is that true disgorgement (stripping the defendant of every penny of an ill-gotten gain)[6] is available *only* when the defendant has breached an obligation of 'good faith or loyalty'.[7] These obligations form a class which is conceptually distinct from obligations arising in contract, tort or unjust enrichment.

* Department of Law, London School of Economics

I would like to thank Peter Birks, Michael Bryan, Gareth Jones and Ewan McKendrick, and participants at the International Conference on the Law of Restitution, University of Tel Aviv, May 1998, for their thought-provoking comments on an earlier draft of this paper

1 'Disgorgement', rather than 'restitutionary damages', is used so as to better differentiate between this remedy and the remedy of restitution for subtractive unjust enrichment ('restitution'): see L. Smith, 'The Province of the Law of Restitution' (1992) 71 *Can Bar Rev* 672, 683–694.

2 Law Commission Report (No 247): *Aggravated, Exemplary and Restitutionary Damages* (1997), recommendations 7–9 and Draft Bill, clause 12, suggesting that the common law ought to be allowed to develop unimpeded, but that, in addition, restitutionary damages (their preferred term) ought to be available where the defendant's wrong (other than a breach of contract) was committed with 'deliberate and outrageous disregard of the plaintiff's rights'

3 *A-G* v *Blake* [1998] 1 All ER 833, 844–846 (CA), suggesting (obiter and without argument or elaboration) that disgorgement for breach of contract might be appropriate where the breach consisted in doing exactly what the contract expressly prohibited or in delivering shortfall performance.

4 See, eg, G. Jones, *Goff & Jones. The Law of Restitution* (London: Sweet & Maxwell, 4th ed, 1993) ('Goff & Jones') 714–734; P. Birks, *An Introduction to the Law of Restitution* (Oxford: Clarendon Press, rev ed, 1989) ('Birks') 326–327 (provided the defendant's breach is cynical); Sir William Goodhart QC, 'Restitutionary Damages for Breach of Contract' [1995] RLR 3; P. Jaffey, 'Restitutionary Damages and Disgorgement' [1995] RLR 30, 40 and 'Disgorgement and Confiscation' [1996] RLR 92, 93, 96; H. McGregor, 'Restitutionary Damages' in P. Birks (ed), *Wrongs and Remedies in the Twenty-First Century* (Oxford: Clarendon Press, 1996) 209 (but only if the defendant's breach is deliberate, cynical and with a view to profit).

5 See ns 2–4, above. Goff & Jones, 721, come close to this, suggesting that the remedy ought to be available whenever the profit could not have been generated 'but for' the breach.

6 'Ill-gotten' imposes an important limiting qualification: the remedy of disgorgement strips only those gains derived as a result of the proven breach of duty.

7 The terminology is not elegant, but it will suffice. As is made clear later, the expression is intended to have a wider compass than fiduciary duties and equitable duties of confidence.

© The Modern Law Review Limited 1999 (MLR 62 2, March) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

Within this class, the disgorgement remedy is independent of the moral culpability of the defendant. Outside this class, disgorgement is not available. It is not an alternative remedial option (even in a limited circumstances) to a claim in tort, contract, or subtractive unjust enrichment. Of course, the same facts may allow a plaintiff to base claims on different causes of action; in this way, one scenario may admit the possibility of both disgorgement and damages claims, for example, but only if claimed as alternative (or perhaps cumulative?) remedies for the breach of distinct obligations.[8] The second claim is that the law of subtractive unjust enrichment is capable of – and, moreover, is the only[9] appropriate restitutionary vehicle for – dealing with a defendant's unauthorised *use* of the plaintiff's property.[10] In itself this is not a new idea.[11] However, its corollary, that the restitutionary remedy must be quantified to reflect the defendant's enrichment *at the plaintiff's expense*, has not been rigorously insisted upon. If this is recognised, then the restitutionary remedy is limited to 'use value'; the claim will not deliver disgorgement. Both of these claims are such that they readily suggest avenues for rational development of the law in the future; this article looks briefly at the possibilities. Before attempting to defend these two claims, it is useful to summarise the existing state of play.

'Disgorgement for wrongs' as a parasitic category

The area of law dealing with 'disgorgement'[12] has been claimed by restitution lawyers.[13] Profits disgorgement is seen as a remedy underpinned by the principle against unjust enrichment, but in a way which is analytically distinct from

8 *United Australia Ltd* v *Barclays Bank Ltd* [1941] AC 1 illustrates this possibility. The plaintiff was the victim when a cheque was fraudulently indorsed to a third party payee. The plaintiff had a claim in subtractive unjust enrichment against the recipient of the proceeds of the cheque *and* a claim in tort against the collecting bank for conversion. The case has been variously interpreted, but it seems to illustrate the potential for alternative *causes of action* on the same facts, rather than the potential for different remedial *responses* to one tort claim. This is all the more evident where, as here, the claims are against different defendants. But cf Birks 316, although also see 321.

9 The distinction between dependent and independent claims can have significant practical consequences, so it is important that the courts choose the appropriate approach: see E. McKendrick, 'Restitution and the Misuse of Chattels – The Need for a Principled Approach' in N. Palmer and E. McKendrick (eds), *Interests in Goods* (London: Lloyds of London Press, 2nd ed, 1998) 914–915.

10 Of course, the facts may also leave open a claim in contract or tort for expectation or compensatory damages.

11 See J. Beatson, *The Use and Abuse of Unjust Enrichment* (Oxford: Clarendon Press, 1991) ch 8; and D. Friedmann, 'Restitution for Wrongs: The Basis of Liability' in W.R. Cornish et al (eds), *Restitution Past, Present and Future* (Oxford: Hart Publishing, 1998) ch 9. However, neither commentator notes the need to limit the remedy. Most other restitution lawyers would categorise these 'wrongful use' claims as dependent on the wrong, not as independent claims in subtractive unjust enrichment.

12 Also termed 'restitutionary damages' or 'restitution for wrongs'.

13 See especially D. Friedmann, 'Restitution of Benefits Obtained Through the Appropriation of Property or the Commission of a Wrong' (1980) 80 *Col LR* 504; R. Sharpe and S.M. Waddams, 'Damages for Lost Opportunity to Bargain' (1982) 2 OJLS 290; G. Jones, 'The Recovery of Benefits Gained from a Breach of Contract' (1983) 99 LQR 443; I.M. Jackman, 'Restitution for Wrongs' [1989] CLJ 302; D. Friedmann, 'The Efficient Breach Fallacy' (1989) 18 *J Leg Studies* 1; J. Beatson, n 11 above; R. O'Dair, 'Restitutionary Damages for Breach of Contract and the Theory of Efficient Breach: Some Reflections' (1993) 46 CLP 113; E McKendrick, n 9 above ch 35; L.D. Smith, 'Disgorgement of the Profits of Breach of Contract: Property, Contract and "Efficient Breach"' (1994) 24 *Can Bus LJ* 121; J. Stapleton, 'A New "Seascape" for Obligations. Reclassification on the Basis of Measure of Damages' in P. Birks (ed), *The Classification of Obligations* (Oxford: Clarendon Press, 1997) ch 8.

© The Modern Law Review Limited 1999

restitution for subtractive unjust enrichment.[14] The distinction is important. The principle against unjust enrichment asserts that a person is not permitted to be unjustly enriched at the expense of another. A claim in subtractive unjust enrichment is an autonomous or independent claim. It depends upon proof that the defendant was (i) enriched (ii) unjustly (iii) at the plaintiff's expense. Existing case law imposes limitations which must be met in proving each of these factors. In the absence of defences, a defendant is then required to pay over the value of the unjust enrichment to the plaintiff. The remedy is termed 'restitution'.[15] 'Disgorgement', on the other hand, is seen as an alternative remedial response to certain wrongs (and not all wrongs fit the bill). The remedy is thus parasitic or dependent on the wrong. In this parasitic category, the principle against unjust enrichment comes into play because, given a pertinent wrong, the defendant's ill-gotten gains are seen as made 'at the expense of the plaintiff' in the sense of 'by doing wrong to the plaintiff'. The most important consequence of this dependent or parasitic analysis is that the defendant's unjust enrichment (or ill-gotten gain) need not correspond to any subtraction from (or loss experienced by) the plaintiff. In this sense the remedy can deliver a windfall to the plaintiff.

At both a theoretical and a practical level there are problems in defending this 'unjust enrichment principle' approach to disgorgement. At a theoretical level, if disgorgement of the profits of wrongdoing is to fall within the principle against unjust enrichment, then the fit must be a little forced. It must be accepted that the particular wrongdoing is itself sufficient reason to class the enrichment as unjust *and* as being gained at the plaintiff's expense. The first is not as difficult as the second.

With the first assertion (ie the notion that the particular wrongdoing is itself sufficient reason to class the enrichment as unjust), it is necessary to distinguish carefully between law and morality. Nevertheless, it could certainly be argued that enrichments gained by certain wrongs (or even all wrongs) are unjust and require reversal by law. Decided cases could then be examined to compile a list of the relevant wrongs, just as, in the area of subtractive unjust enrichment, there is a list of 'unjust factors'. But it is the *justification* for including certain wrongs and excluding others that causes so many difficulties. It is not difficult to see why. The disgorgement remedy, if it is parasitic, supplies an alternative remedial response: with certain wrongs the plaintiff can choose either disgorgement or damages as a remedy. This choice needs compelling justification. It is not a choice between alternative causes of action; it is a choice between alternative remedial strings for the one cause of action. To justify this necessarily involves some rethinking the law of obligations (or certain obligations)[16] so that they are recognised as directed not only at preserving the plaintiff from harm,[17] but also at 'punishing' the defaulting defendant. The 'punishment' is not necessarily economic;[18] rather the disgorgement remedy does – and is intended to – coerce or discipline the defendant into complying with the underlying obligation. Such attempts to compel 'good' behaviour[19] may be laudable, but they have not traditionally been seen as the function of obligations in contract or tort. Equitable obligations of good faith and

14 Not all commentators adopt this approach, but this describes the majority view.

15 Or, in full, 'restitution for subtractive unjust enrichment'.

16 Since it is conceded that not all wrongs require disgorgement.

17 Intended widely, as harm to the plaintiff's *status quo*, or harm to the plaintiff's legitimate expectations.

18 Especially since the defendant is in no worse a position than if the obligation had not been breached. This is not necessarily the case with expectation or compensatory damages.

19 cf O.W. Holmes, 'The Path of the Law' (1897) 10 *Harvard Law Rev* 458.

© The Modern Law Review Limited 1999

loyalty are different: there the orthodox remedial response is to punish the defendant regardless of harm to the plaintiff; the obligations are seen as compelling the defendant to display more than 'the morals of the market place'.[20]

The second assertion, that the enrichment (or the ill-gotten gain) is gained *at the plaintiff's expense*, creates even more theoretical difficulties. It is simply conclusionary to say that, since the gains are made by doing a wrong to the plaintiff, they are therefore made at the plaintiff's expense. The statement itself is not intuitively correct. Indeed, if the defendant's obligation is simply not to cause harm or to meet expectations, then the 'expense' suffered by the plaintiff because of a breach is more naturally seen as the financial loss caused by the harm or the unmet expectations, losses which are already remedied in tort and contract. If the 'at the plaintiff's expense' requirement is simply ignored, then all that is left is the (admittedly beguiling) assertion that individuals should not be allowed to profit from their own wrongs. Enforcement of this principle might be seen as a desirable legal objective, but it cannot yet be seen as a general and unqualified rule of the existing law of contract, tort, or even unjust enrichment as that last principle is rigorously defined.

Even at a practical level, there are problems in defending this 'parasitic claim' approach to disgorgement. The judges themselves rarely adopt an overtly restitutionary analysis.[21] It is therefore left to commentators to decide whether there *is* a parasitic category of 'restitution for wrongs' and, if so, which cases ought to be assigned to it. Confident assertion of the appropriate classification of a particular case is difficult: the quantum of a remedy may be equally consistent with disgorgement, restitution, or either compensatory or expectation damages.[22] In fact, if disgorgement is truly parasitic, a plaintiff will invariably be able to claim compensatory or expectation damages as an alternative remedy, and may even be able to claim restitution (based on an alternative and independent cause of action). Often the cases do not make it clear which route is being pursued. Even accounting for this difficulty, it is clear that not all wrongs warrant the remedy of disgorgement. Disgorgement (even if defined widely to include 'use value' or 'expenses saved', contrary to the terminology adopted in this article) is only available for certain equitable wrongs[23] and perhaps also for some proprietary torts;[24] otherwise the remedy is denied.[25] Various theories have been advanced to

20 *Meinhard* v *Salmon* 164 NE 545 (1928), 546 *per* Cardozo CJ.

21 Reflecting the fact that the principle against unjust enrichment has only recently received its judicial imprimatur.

22 The terminology becomes unwieldy very quickly. As far as possible, this article uses the terms 'compensatory damages' to refer to the normal remedy for commission of a tort, 'expectation damages' to refer to the normal remedy for breach of contract (both of these being plaintiff-oriented loss-based remedies), 'restitution' to refer to the normal remedy for subtractive unjust enrichment, and 'disgorgement' to refer to the remedy requiring payment over of ill-gotten gains. Notably only the last is described without reference to a causative event. Part of the thesis advanced here is that disgorgement is the normal remedy for breach of obligations of loyalty and good faith. Another point merits notice: the first two remedies are inevitably termed 'damages', the last two are not. This is significant. The first two are inevitably personal remedies requiring payment of a sum of money. The last two may be remedies ordered in this form, but they may also be proprietary remedies compelling transfer of assets in specie. This issue is taken up later.

23 eg breach of fiduciary duties and breach of confidence. The term is used loosely here, but its ambit – in the context of this article – is made clear later

24 Although this article argues that there is a crucial distinction between 'use value' and 'disgorgement' remedies.

25 See *Halifax Building Society* v *Thomas* [1996] 2 WLR 63, [1995] 4 All ER 673, 680 *per* Peter Gibson LJ; although also note n 2 and 3 above. *Statutory* disgorgement remedies do exist (principally for intellectual property torts), but these remedies are not necessarily motivated by the same imperatives which underpin disgorgement at common law or in equity. It follows that these statutory examples

© The Modern Law Review Limited 1999

explain and justify the limitations, but none seems to be both consonant with decided cases and sufficiently precise to provide a predictive tool.[26] To complicate matters still further, the equitable wrongs appear to require full disgorgement of all ill-gotten gains, while the proprietary torts require only disgorgement of the 'use value' of the property to the defendant. In both cases this is regardless of whether the wrongdoing is innocent or cynical. In short, the picture is confused. These practical incidents make confident assertion of the theoretical underpinnings of the area very difficult. With that introduction, it is possible to address the specific issues which are central to this article.

Classifying and reclassifying proprietary torts

The argument presented in this section is that the law of subtractive unjust enrichment is capable of – and is, in fact, the only appropriate restitutionary vehicle for – dealing with a defendant's unauthorised *use* of the plaintiff's property.[27] On this approach, the defendant will make 'restitution' by paying over to the plaintiff the 'use value' of the misused asset (no more and no less). There are three distinct issues here, although not all are given equal attention. The first is the idea that subtractive unjust enrichment is *capable* of dealing with such claims. This is not new; the detailed arguments have already been put by Beatson (although his conclusions do not match those advanced here).[28] Here greater emphasis is given to the corollary to this, that the restitutionary remedy must be quantified to reflect the defendant's enrichment *at the plaintiff's expense*. This constraint has somehow been ignored. Its recognition makes it plain that the restitutionary remedy must be limited to 'use value'; a subtractive unjust enrichment claim will not deliver disgorgement. Finally, there is the idea that the autonomous claim is the *only* restitutionary vehicle in these misuse of property cases; 'restitution for wrongs', parasitic on the tort, is not available as an independent option. Admittedly this assertion is impossible to prove affirmatively. However, it does appear to be true that there are no common law misuse of property cases where the only *possible* explanation is a parasitic claim of 'restitution for wrongs'.[29] Where full disgorgement of all ill-gotten gains is ordered, it always seems arguable that the defendant has breached some concurrent equitable obligation owed to the plaintiff.[30] Such a

may not illustrate – never mind justify – conclusions concerning a common law practice or policy governing disgorgement. The focus in this article is exclusively on the position at common law (taken to include equity).

26 Notwithstanding that, the ideas advanced here owe much to those analyses; see especially n 13 above.

27 Of course, non-restitutionary claims are possible based on alternative causes of action.

28 J. Beatson, n 11 above. Also see D Friedmann, n 11 above. Both commentators advocate a much wider view than that advanced here: they assert that the independent subtractive unjust enrichment claim is available even when the defendant has 'invaded' the plaintiff's *non*-proprietary interests; moreover, neither commentator notes the need to limit the quantum of the remedy. The more general notion of *alternative* claims for unjust enrichment and for wrongs is widely conceded, but most restitution lawyers would put misuse (as opposed to receipt) of property cases exclusively in the latter category, not seeing the former as capable of dealing with the issue.

29 See J. Beatson, n 11 above. But not all would agree; see, eg, P. Birks, 'Restitution and Wrongs' (1982) 35 CLP 53, 63–65. However, Birks' exceptions (apart from conversion – see n 77) are examples which might be alternatively classified as breaches of equitable obligations.

30 For the common law cases on intellectual property torts, this assertion is less overtly secure. Consider passing off. On the arguments advanced here, *if* this tort is regarded as involving the *use* of another's property (which may be arguable, given the tenor of some judgments), then a claim in subtractive unjust enrichment ought to be possible (and market valuation of the 'use value' might deliver an account of profits – see n 58). True disgorgement, on the other hand, would not be a remedial option on the analysis advanced here, at least not until 'good faith' (or 'absence of bad faith') was regarded

© The Modern Law Review Limited 1999

breach gives rise to an independent cause of action with the remedy of disgorgement; in this class, disgorgement is not a second remedial string parasitic on a claim which would ordinarily give rise to some alternative remedy.[31] These equitable obligations are considered in the final section of this article.

The discussion in this area is deliberately restricted to cases delivering common law rather than statutory remedies for proprietary torts.[32] Within this category, separate consideration is given to cases where the defendant has wrongfully used the plaintiff's property (without destroying it) and cases where the defendant has wrongfully converted or made investments with the plaintiff's property (so that the original asset is no longer in the defendant's possession). The aim is to articulate a framework which explains and justifies the results of decided cases, rather than one which unites their judicial reasoning. The proponents of restitution for wrongs as a parasitic category adopt the same strategy. There is no other route: the language of restitution is rarely found in the older cases, notwithstanding that it might now be the most appropriate rationalisation.

Unauthorised use of a plaintiff's property: an unjust enrichment analysis

Unauthorised use of another's property is invariably a tort. The plaintiff victim can sue for compensatory damages. But can the plaintiff demand disgorgement as an alternative remedy for the tort? Or can the plaintiff rely on an alternative cause of action, and sue in subtractive unjust enrichment to obtain restitution? With several notable exceptions, most of the cases classed by restitution lawyers as illustrating disgorgement for unauthorised use of the plaintiff's property are couched in the language of compensatory damages to remedy the tort. Sometimes this is clearly the appropriate measure; but sometimes the judicial language conceals conceptual obstacles which demand alternative justifications. These alternative justifications lead into the disputed area of 'disgorgement' and 'restitution' remedies.

An example illustrates the issues. In *Strand Electric & Engineering Co Ltd* v *Brisford Entertainments Ltd*[33] the defendant wrongly kept and used the plaintiff's theatre equipment after the conclusion of a period of hire. The plaintiff was held entitled to compensatory damages calculated as a reasonable hire charge for the period of detention.[34] Since the plaintiff was in the business of hiring out the equipment, this approach is easily justified. Compensatory damages are designed to put the plaintiff in the same financial position as if the tort had not been committed. The measure of damage is lost hiring fees. Normally a plaintiff would have to allow for the chance that the goods could not be hired out commercially. However, where the defendant is clearly a putative hirer, it hardly lies in the defendant's mouth to claim that there would have been lay-times during the relevant period.

But this traditional tort analysis will not meet all the fact situations. If the plaintiff is not in the business of hiring out equipment, then the financial loss

as demanded at law. This possibility is considered later. Some passing off cases, in distinguishing between deliberate and innocent infringement, may lend limited support to a 'good faith' argument. More detailed consideration of these specific torts is clearly warranted, but is beyond the scope of this article.

31 Contrast this with claims for 'restitution for wrongs' alleged to be parasitic on a tort or breach of contract.

32 See n 25 above.

33 [1952] 2 QB 246 (CA).

34 Although Denning LJ, *ibid* 254–255, also viewed the claim as resembling an action for restitution.

© The Modern Law Review Limited 1999

resulting from the defendant's wrongful detention of the goods may be nil. Judicial reliance on compensation based on notionally lost hiring charges or licence fees will then not withstand scrutiny. Compensatory damages are designed to put the plaintiff in the position he or she would have been in if the tort had not been committed. This means the position, so far as money can do it, that the plaintiff would have been in had the goods been at his or her disposal for the relevant period. If the goods would not have been gainfully used by the plaintiff, then the loss (absent amenity damages) is nil.[35] The function of tort law is not to imply hire bargains between the parties; this is so whether or not the plaintiff might have been willing to hire out the goods.[36]

Nevertheless, a sense of justice demands that the plaintiff be given a remedy – and in many cases the judges have acceded to these demands, albeit on questionable grounds. Commentators often claim these cases as examples of restitution for wrongs (ie as cases where the defendant is required to disgorge ill-gotten gains regardless of any loss to the plaintiff). The analytical difficulties in complying with orthodox compensation measurements are then replaced by difficulties of principle. Which wrongs will justify disgorgement? When will disgorgement capture all the defendant's gains? And when will it be restricted to saved expenditure (measured by notional licence or hire fees)?

Again an illustration highlights the issues. In *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd*,[37] the defendant built fourteen houses without seeking the plaintiff's approval and so in breach of a restrictive covenant which had been registered as a land charge. The judgment indicates that the breach did not reduce the value of the plaintiff's land, but it did allow the defendant to make a profit of £50,000 and to save, in the judge's estimate, £2,500 in declining to bargain with the plaintiff for a release from the covenant. Brightman J decided that the plaintiff ought to receive £2,500 as compensation for the wrong.[38] The difficulties are clear. If the remedy is truly compensation, then £2,500 is too much: the plaintiff's proven financial loss, measured by the reduced value of its land, is nil.[39] On the other hand, if the remedy is truly disgorgement for wrongs, then £2,500 is not enough: the ill-gotten gains amount to £50,000, not £2,500.

---

35 cf R.J. Sharpe and S.M. Waddams, 'Damages for Lost Opportunity to Bargain' (1982) OJLS 290.

36 Some commentators reject the compensatory damages approach only where it is clear that the plaintiff *would not* have hired out the goods (or allowed the trespass, or released the restrictive covenant). See, eg, Law Commission, *Aggravated, Exemplary and Restitutionary Damages* Consultation Paper No 132 (1993) para 7.10; H. McGregor, n 4 above, 210; P. Jaffey, 'Restitutionary Damages and Disgorgement' [1995] RLR 30, 33. But the plaintiff's willingness to accept money is irrelevant; damages are calculated to remedy financial harm, regardless of whether the plaintiff would have agreed to suffer the harm in exchange for money. It follows that the compensatory damages approach is inappropriate in *all* cases where the plaintiff has suffered no financial harm. see eg, *Surrey CC* v *Bredero Homes Ltd* [1993] 1 WLR 1361, 1369 *per* Steyn LJ.

37 [1974] 1 WLR 798.

38 Under Lord Cairns' Act, in lieu of an injunction requiring demolition of the houses. Probably too little attention is given to the role of this statute, especially when cases are used to support generalisations. However, the valuations in this case do provide a pointed illustration of the potential impact of alternative legal responses.

39 Brightman J held that the plaintiff could be put in the position it would have been in if the covenant had not been breached *either* by a payment that would put the plaintiff in the same financial position it would have been in had the houses not been built (ie nil) *or* by a payment that would put the plaintiff in the same position it would have been in if the defendant had bargained for a release from the covenant so that the houses did not constitute a breach of covenant (ie £2,500). As noted earlier, the latter route effectively substitutes an implied bargain between the parties for the granting of permission to commit torts or breaches of contract; this is something quite outside orthodox tort or contract law principles (although perhaps it can be justified under Lord Cairns' Act).

© The Modern Law Review Limited 1999

This last assertion needs elaboration. The cases concerning defaulting fiduciaries provide the clearest instruction in the process of quantifying true disgorgement remedies. In the fiduciary cases the obligation to disgorge is non-controversial. Nevertheless, there is a causation test: only those gains derived from the breach need to be disgorged. The distinction between such gains and others in the hands of the defaulting fiduciary is not always easy. Sometimes fiduciaries have to disgorge all the profits of an activity;[40] at other times they must disgorge only part of the total profit derived;[41] and possibly there are times when fiduciaries may keep the profits subject to disgorgement of a fee regarded as sufficiently representative of the gains attributable to the breach of duty.[42] The particular facts of the case indicate the appropriate method of quantification; there is no hard and fast rule. The same must be true of disgorgement remedies operating outside the arena of fiduciary breaches.

Returning to *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd*,[43] it is clear that if the defendant's activity had constituted a breach of *fiduciary* duty, the remedy would have required disgorgement of £50,000. This suggests that the term 'disgorgement' does not properly characterise the remedy awarded. This is all the more so with cases where the defendant's profits are not even raised as a relevant issue; instead, the focus is entirely on assessing the 'use value' to the defendant of the plaintiff's property.[44] The suggestion advanced here is that the remedy in all these cases is better seen as restitution for subtractive unjust enrichment rather than disgorgement for wrongs, although this does require some extension of orthodox unjust enrichment analysis.

To fit these cases within the accepted subtractive unjust enrichment framework, the defendant must have been (i) enriched (ii) unjustly (iii) at the plaintiff's expense. In the absence of any defences, the defendant is then required to pay over the value of the enrichment to the plaintiff. In these cases where the defendant has made unauthorised use of the plaintiff's property, the fact that the defendant has been enriched is easily demonstrated; so too that the enrichment is unjust, being without the plaintiff's consent (a paramount 'unjust factor').[45] Despite its intuitive attraction, the difficulty is to demonstrate that the enrichment is 'at the plaintiff's expense' notwithstanding that the plaintiff has suffered no financial loss.

The problem with the requirement that the enrichment must be at the plaintiff's expense is that it is so easy to slip from that formulation into a formulation which requires the defendant's (financial) gain to be matched by the plaintiff's (financial)

40 Perhaps subject to an allowance which appropriately reflects the fiduciary's contribution to the choice and operation of the profit-making venture: *Boardman* v *Phipps* [1967] 2 AC 46.

41 *Warman International Ltd* v *Dwyer* (1995) 182 CLR 554; *Scott* v *Scott* (1963) 109 CLR 649, 662; *Paul Davies (Aust) Pty Ltd (in liq)* v *Davies* [1983] 1 NSWLR 440; *Australian Postal Corpn* v *Lutac* (1991) 21 NSWLR 584 Also see L. Aitken, 'Loan Funds and Trustee's Profit' (1993) 13 *Legal Studies* 371.

42 It is true that the facts are unlikely to suggest a breach and yet find the profits not attributable, at least in part, to the breach. However, (and in another context) the controversial decision in *Re Tilley's WT* [1967] 1 Ch 1179 suggests the possibility.

43 n 37 above.

44 See n 58 below.

45 The issue of the 'unjust factor' deserves greater elaboration than is possible here. Suffice it to note that the lack of consent may be evidenced positively (as in *Strand Electric & Engineering Co* v *Brisford Entertainments* [1952] 2 QB 246 – although this is a compensatory damages case, not a subtractive unjust enrichment case – where the defendant was notified by the plaintiff) or negatively (as appeared to be the case in *Whitwham* v *Westminster Brymbo Coal and Coke Co* [1896] 2 Ch 538, where the plaintiff was ignorant of the use). The latter is more likely in cases where restitution, not compensatory damages, is claimed Nevertheless, either scenario is possible; 'lack of consent' is therefore preferred to 'ignorance' to describe the 'unjust factor'.

© The Modern Law Review Limited 1999

loss.[46] The slip is encouraged because this alternative formulation appears – incorrectly – to underpin the remedy when the defendant has received some physical asset from the plaintiff. Then, it is true, the defendant's financial gain is carefully assessed; subjective devaluation is allowed. But the plaintiff's corresponding financial loss is *not* so carefully considered; assessment of the unjust enrichment is never reduced upon proof that the *plaintiff* does not value the subtraction as highly as the defendant does. In fact, 'at the plaintiff's expense' merely requires proof that the defendant's enrichment is derived directly and exclusively *from* the plaintiff (in the form of assets or services); this is all that the 'subtractive' element requires.[47] The cases involving provision of services by the plaintiff show that this must be the case.[48] The statement imposes a relational restriction, not a financial one. Given that relational factor, the defendant's enrichment is initially assessed at its market value (*quantum meruit* or *quantum valebat*), perhaps then reduced if the facts suggest that subjective devaluation is appropriate. The reduction is allowed because the gist of the remedy is to strip enrichments, not to return either plaintiff or defendant to their *status quo ante*. These steps are routinely adopted when the defendant has received goods or services from the plaintiff in circumstances which suggest that the receipt confers an unjust enrichment.[49] The analogy between enrichments gleaned from use of the plaintiff's own skills or labour and those gleaned from use of the plaintiff's own assets seems simple and compelling.[50] Yet the former are routinely consigned to the category of subtractive unjust enrichment and the latter to unjust enrichment by wrongdoing. The distinction does not seem defensible.

This is all the more so when the jurisprudential underpinnings of the law of unjust enrichment are considered. The gist of the action is not to remedy loss, nor to right wrongs. It is to compel the defendant to give up unjustified enrichments. The subtractive element in the cause of action serves to determine the person to whom the enrichment must be given up. Nevertheless, it is commonly said that the plaintiff cannot recover more than he or she has lost.[51] But the true import of that statement is easily illustrated. Assume that a plaintiff cabinet-maker does substantial work on the defendant's kitchen. Assume, further, that there is no contract providing for payment (the reasons are immaterial). It might reasonably be the case that the plaintiff's labour had a market value of £5,000, but that

46 See, eg, L.D. Smith, 'Three-Party Restitution: A Critique of Birks's Theory of Interceptive Subtraction' (1991) 11 OJLS 481, who sees 'at the expense of' as a 'zero sum gain' (ie 'your plus is my minus'), likewise the Canadian view that an enrichment should be matched by a corresponding deprivation: eg *Air Canada* v *British Columbia* (1989) 59 DLR (4th) 161 (Can), 193–194 *per* La Forest J. And restitution is then seen as returning the parties to their *status quo ante*: see, eg, Birks 132; A. Burrows, *The Law of Restitution* (London Butterworths, 1993) ('Burrows') 18. But this is not the *aim* of the unjust enrichment principle. Its aim is to reverse the unjust enrichment. Neither party is necessarily returned to the position occupied before the enriching event: see J. Stapleton, n 13 above, 197–198.

47 See Sir Peter Millett, 'Restitution and Constructive Trusts' (1998) 114 LQR 399, 409.

48 It is never to the point that the plaintiff might otherwise have been unemployed, and so has suffered no financial loss (other than out of pocket expenses).

49 These are difficult cases, but see eg, *BP Exploration Co (Libya) Ltd* v *Hunt (No 2)* [1979] 1 WLR 783, aff'd [1983] 2 AC 352, *Greenwood* v *Bennett* [1973] QB 195; *Deglman* v *Guaranty Trust Co of Canada* [1954] SCR 725, [1954] 3 DLR 785 (Can).

50 And is clearly recognised in Birks, at 129, notwithstanding his preference for a 'restitution for wrongs' approach. The analogy is not entirely unexpected: rights to personal autonomy and rights in property are both far more complicated 'bundles of rights' than other rights (such as rights under a contract) and their various modes of protection have obvious parallels.

51 The inference being, again, that *financial* gain must be matched by *financial* loss: see, eg, Birks 313; A. Burrows and E. McKendrick, *Cases and Materials on the Law of Restitution* (Oxford: Clarendon Press, 1997) 569.

 © The Modern Law Review Limited 1999

installation of a modern kitchen increased the value of the defendant's house by £10,000. It might also be the case that the plaintiff did the work on days when he could not or would not otherwise have worked. If the plaintiff has no contractual claim for payment for his work, the law of unjust enrichment can compel the defendant to make some payment.[52] Moreover, it is uncontroversial that the defendant's 'enrichment at the plaintiff's expense' will be seen as £5,000,[53] not as £10,000 (even though this is the defendant's improved financial position as a result of the plaintiff's work), nor as nil (even though this is the plaintiff's financial loss, if we assume that the plaintiff could not or would not have gainfully used the time in alternative labour). The parallels with cases where a defendant has made unauthorised use of the plaintiff's assets are stark.

The case of *Ministry of Defence* v *Ashman*[54] provides another illustration. The defendant tenant was a trespasser occupying the plaintiff's premises. The premises were normally let at a concessionary rent of £95 per month to service personnel, although their market rental value was £472 per month. The defendant wanted local authority housing (apparently priced at £145 per month). The Court of Appeal decided that the plaintiff was entitled to restitutionary damages assessed at the value of local authority housing. Restitution lawyers normally consign the case to the category of 'restitution/disgorgement for wrongs' because the remedy is not related to the plaintiff's financial loss (ie any lost concessionary rent, which would have been the measure of *compensatory* damages for the trespass). It is related only to the defendant's 'enrichment'. However, on another view it is clear that all of the defendant's enrichment was *from* the plaintiff, and in that sense was 'at the plaintiff's expense'. It is a separate and more difficult issue – for both analyses – to determine whether the market value of that enrichment should have been subjectively devalued in favour of a conscious wrongdoer simply because the defendant *desired* cheaper accommodation (even though the facts suggest none was available to her at the time).

In this context, it is important to recognise that where a defendant makes unauthorised use of a plaintiff's asset, the mere fact that the asset is not 'profit-earning' in the plaintiff's hands (or is, but only to a limited extent) ought to be immaterial in calculating the *defendant's* enrichment. The enrichment calculation requires only that a 'use value' can be quantified.[55] The profit-earning nature of the asset (in the plaintiff's hands) ought to be relevant only in assessing compensatory damages.

In consigning these proprietary tort cases to the category of subtractive unjust enrichment, the historical parallels between judicial treatment of cases where the defendant has had the benefit of the plaintiff's services and cases where the defendant has had the use of the plaintiff's goods are also significant. The 'implied contract' approach adopted in the former cases has been justifiably criticised and is now overthrown. But the very same approach is still routinely adopted in cases of trespass to land and goods: as an alternative to the compensation claim for the tort, the plaintiff is seen as entitled to *mesne* profits or notional hire charges or licence fees. These remedies all import a fictional agreement between the parties to pay for the unauthorised use of the plaintiff's assets. They do this because common sense

52 See, eg, *Pavey & Matthews Pty Ltd* v *Paul* (1986) 162 CLR 217 (Aus).

53 Sometimes this sum may be reduced because of subjective devaluation, although see *Rover International Ltd* v *Cannon Film Sales Ltd* [1989] 1 WLR 912

54 (1993) 66 P & CR 195 (although Lloyd LJ expressly rejected a restitutionary analysis).

55 But see *Strand Electric & Engineering Co Ltd* v *Brisford Entertainments Ltd* [1952] 2 QB 246, 252, 257, where Romer and Somervell LJJ leave this issue open.

© The Modern Law Review Limited 1999

suggests that any other reaction would clearly result in the defendant being unjustly enriched. They do not do this because the defendant's particular wrong is so heinous as to warrant punishment by compelling disgorgement of the ill-gotten gains. This is not the tenor of the judgments nor the measure of the intervention ordered. Recognition of the unjust enrichment basis of these fictional agreements is slow in coming, but there are signs of judicial awakening. In *Ministry of Defence* v *Ashman*,[56] Hoffmann LJ pronounced that it was now time 'to call a spade a spade' and recognise that an award of *mesne* profits for trespass was a restitutionary remedy.[57]

Up to this point the argument has been devoted to advancing the idea that 'use value' remedies awarded in proprietary tort cases might reasonably be regarded as restitution for subtractive unjust enrichment. If this were accepted, then an independent claim in unjust enrichment would simply be an alternative to the claim in tort. It is also part of the thesis advanced here that the *tort* claim does not give rise to alternative remedial responses: more specifically, the remedy of disgorgement is not available as an alternative to the remedy of compensation. Proof of this assertion is impossible. Nevertheless, the relevant cases invariably adopt a notional 'consent to use' fee in quantifying the remedy, and do this regardless of the defendant's actual ill-gotten profit. This accords more naturally with an assessment of the defendant's subtractive unjust enrichment at the plaintiff's expense, rather than with disgorgement of the defendant's ill-gotten gains.[58]

If unauthorised use of the plaintiff's assets were seen as giving rise to a restitutionary claim grounded only in subtractive unjust enrichment, would the resulting analysis provide any advantages which are not delivered by an analysis grounded in disgorgement for wrongs? Three important benefits seem to follow. First, the re-classification of the 'use value' cases would highlight the fact that true disgorgement (of *all* ill-gotten gains) is currently restricted to a very narrow class of cases, cases which invariably concern equitable, not common law, obligations.[59] Secondly, the re-analysis would allow for a rational explanation of why the remedy in these re-classified cases is restricted to 'use value', rather than full disgorgement, and why only limited classes of wrongs – involving unauthorised use of property[60] – give rise to this 'restricted disgorgement' remedy. Finally, the existing scholarship on subtractive unjust enrichment would indicate how 'use value' should be calculated in different circumstances.

The last of these advantages merits elaboration. If the 'use value' remedy awarded in proprietary tort cases is seen as being for subtractive unjust enrichment,

56 n 54 above.

57 *ibid* 201 Also see *Strand Electric & Engineering Co Ltd* v *Brisford Entertainments Ltd* [1952] 2 QB 246, 254–255 *per* Denning LJ: 'It is an action against [the defendant] because he has had the benefit of the goods. It resembles therefore an action for restitution *rather than* an action for tort' (emphasis added).

58 See, eg, *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd* [1974] 1 WLR 798; *Penarth Dock Engineering Co Ltd* v *Pounds* [1963] 1 Lloyd's Rep 359; *Whitwham* v *Westminster Brymbo Coal and Coke Co* [1896] 2 Ch 538, especially 541–542 *per* Lindley LJ. In *Strand Electric & Engineering Co Ltd* v *Brisford Entertainments Ltd* [1952] 2 QB 246 it was commented that if the defendant *had* profited from the use, then disgorgement of those profits would *not* be required (*per* Somervell LJ at 252) *or* might be (*per* Denning LJ at 255). On the other hand, there are cases where a percentage of the defendant's profits might be regarded as the only appropriate measure of a 'right to use' fee. This will be particularly so where the use is intimately tied to the profits generated and where, on any objective assessment, a 'use fee' would only ever be negotiated to reflect that fact. *Edwards* v *Lee's Administrators* 96 SW 2d 1028 (1936) (Kentucky) (the Great Onyx Cave case) might be seen as such a case, although this was not the reasoning adopted by the court. Also see the arguments put by Beatson and Friedmann, n 28 above.

59 The reason for this is considered below.

60 Although 'property' may be an expanding concept: see below at nn 93–95 and associated text.

© The Modern Law Review Limited 1999

rather than for the wrong (and dependent on proof of the wrong), then issues of subjective devaluation and change of position become important. They enable proper quantification of the restitutionary remedy when the defendant's tort involves innocent (albeit unauthorised and wrongful) use, or no use, or no profitable use, of the plaintiff's property.

Innocent but unauthorised use of the plaintiff's property will constitute a tort, and the defendant's liability for compensatory damages will be strict. But the same is not necessarily true of the remedy in unjust enrichment. If the defendant reasonably (although incorrectly) believed that the plaintiff had given consent, or was obliged to give consent, to the use, then it seems open to the defendant to argue for subjective devaluation of the use; perhaps even to devalue the use to nil, on the basis that the defendant would not have made use of the asset on terms other than those assumed to exist.[61] The obvious analogy is with the innocent recipient of a mistaken non-monetary benefit.[62]

If the defendant's profitable wrongdoing does not involve *any* use of the plaintiff's property, an analysis based on subtractive unjust enrichment can explain why the defendant should be liable for compensatory damages for the tort, but not for restitution for unjust enrichment.[63] The position of a warehouseman wrongfully detaining goods is commonly cited as an illustration. The *possibility* of use where use was not made is not an enrichment,[64] notwithstanding that the defendant may have committed a wrong (even cynically, and with or without causing financial damage to the plaintiff) and may even have profited from that wrong. *Stoke-on-Trent City Council* v *W & J Wass Ltd*[65] provides an illustration. The defendant conducted an unauthorised market, thereby interfering with the plaintiff's authorised market. No disgorgement was ordered. The reasoning in the case has been justifiably criticised,[66] but the result may be right. The defendant undoubtedly

61 Perhaps this explains *Morris* v *Tarrant* [1971] 2 QB 143, where the defendant husband was, technically, a trespasser as against his ex-wife, but was not required to pay any 'use value', there had been inconclusive and continuing negotiations between the parties during the period. Another difficult case is *Inverugie Investments* v *Hackett* [1995] 1 WLR 713. The parties were in continuing dispute over their respective rights. Assuming the dispute was legitimate, it seems proper that restitution for unjust enrichment might be limited by subjective devaluation. The remedy actually awarded therefore seems more easily justified as compensatory, not restitutionary, notwithstanding Lord Lloyd's view (at 718) that the user principle was not exclusively compensatory or restitutionary. As in *Strand Electric & Engineering Co* v *Brisford Entertainments* [1952] 2 QB 246, discussed earlier, damages for lost rent in *Inverugie* were not reduced to allow for periods of likely vacancy. The case seems more difficult to justify on restitutionary principles, although see C. Mitchell, 'Meane Profits and Restitutionary Damages' [1995] LMCLQ 343.

62 Of course, a reduction might conceivably be allowed even if the analysis were grounded in disgorgement for wrongs. But then another difficulty arises: how to distinguish cases where a reduction is permissible from all those where it is not, the defendant's innocence notwithstanding (as in the fiduciary cases).

63 And, if appropriate, return of the asset itself or its value, a remedy which some see as restitutionary but others do not, preferring instead to regard the non-transfer of title as pre-empting the need for restitution. See P. Birks, 'Property and Unjust Enrichment: Categorical Truths' [1997] NZLR 623 and the literature cited therein.

64 See eg, *Strand Electric & Engineering Co Ltd* v *Brisford Entertainments Ltd* [1952] 2 QB 246, 250, 254, 256 The only 'enrichment' is possession of the asset. This is remedied by compelling return of the asset or its value (see n 63 above). By way of contrast, alternative rationalisations of the proprietary torts cases (see n 13 above) – based on either 'loss of dominion' or 'loss of the opportunity to bargain' – would suggest there ought to be some remedy, restitutionary or compensatory, for mere unauthorised possession. With 'loss of dominion' this is perhaps less obvious, but (on reasoning criticised here) the plaintiff's loss of dominion has a money value; this means that so too does the defendant's gain of dominion; restitution of an enrichment can therefore be justified.

65 [1988] 1 WLR 1406.

66 eg I.M. Jackman, n 13 above, 307; Burrows 392–393; Goff & Jones 722–723.

© The Modern Law Review Limited 1999

caused a nuisance, thus entitling the plaintiff to claim compensation for any resulting loss. However, the defendant did not '*use*' the plaintiff's proprietary market right: that market right did not give the plaintiff any proprietary interest in surrounding sites, in particular the site used by the defendant.[67] Even if the defendant's wrong generated a profit, it was not from the physical use of the plaintiff's property; it was not *from* the plaintiff in this sense.[68] The only way that a subtractive unjust enrichment analysis could get around this, and so capture the profit (or part of it), is to enlarge the concept of property to include all 'rights', including those derived from personal obligations owed to the plaintiff in contract and tort. This seems both undesirable and unwarranted; it would deny any distinction between property and obligation.[69]

Consider one more example. If the defendant's wrongful use of the plaintiff's asset is not profit-generating, an analysis based on subtractive unjust enrichment can explain why, notwithstanding this, it may be legitimate to require restitution from the defendant. Consider a defendant who makes unauthorised use of the plaintiff's car as a taxi, but is unsuccessful in the venture. The defendant's 'enrichment' is the use value of the car. Prima facie this is the market value of such hire, and, assuming that the defendant is aware that the use is unauthorised, there seems to be no basis for subjective devaluation. On the arguments advanced here, this enrichment is derived *from* the plaintiff, and is thus 'at the plaintiff's expense'. It is irrelevant that the plaintiff has suffered no financial loss from the defendant's use: remedying this is the function of tort law, not the law of unjust enrichment.[70] It is also irrelevant that the defendant has not made a financial gain from the wrong,[71] although this would be crucial in any real disgorgement remedy.[72] On the other hand, if the facts indicated that the defendant's use was innocent, then not only is subjective devaluation of the benefit a possibility, but so is a complete change of position defence: it is conceivable that the defendant would not have used the car at all had the true facts been known. Significantly, innocence would be irrelevant in assessing the defendant's liability in tort.

Unauthorised use of a plaintiff's property: applying the analysis to investments made using the plaintiff's property

One further matter merits attention in this context. The previous section was devoted exclusively to cases where the defendant wrongfully used the plaintiff's property without destroying it or passing it to some third party. But, on a very wide

67 The case is thus fundamentally different from that of *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd* discussed earlier, where the plaintiff did have a proprietary interest in the land (mis)-used by the defendant. Classification of certain interests as proprietary and others as personal is controversial, especially at the margins, but as the law currently stands these cases fall on opposite sides of the divide, notwithstanding superficial similarities.

68 Although, as already noted, the proponents of 'disgorgement for wrongs' give this expression an additional meaning – a legal meaning not tightly related to the linguistic sense – that the term also connotes 'from a wrong to the plaintiff'.

69 However, for a wide view of the rights warranting such protection, see LM. Jackman, 'Restitution for Wrongs' [1989] CLJ 302; D. Friedmann, 'The Efficient Breach Fallacy' (1989) 18 *J Leg Studies* 1.

70 And if the plaintiff *has* suffered a loss (eg if the plaintiff normally used the car as a taxi), then that loss is appropriately remedied by compensatory damages for the tort.

71 Note the analogy with mistaken payments: ignoring possible defences, the law of subtractive unjust enrichment will require a mistaken payee to make restitution notwithstanding subsequent loss or dissipation of the mistaken payment.

72 As fiduciary law indicates: eg, if the fiduciary engages wrongfully but *unprofitably* in a competing business, the principal will have no remedy for the breach. Of course, given amenable facts, other causes of action (in contract, tort or unjust enrichment) may provide the plaintiff with a remedy.

© The Modern Law Review Limited 1999

definition, 'use' might also include instances where the defendant wrongfully converts or makes investments with the plaintiff's property so that the original asset is no longer in the defendant's possession. In these latter cases, full disgorgement is sometimes claimed through the back door, invariably without articulating what is being done, perhaps without recognising what is being done. The issue arises where the plaintiff has a proprietary claim to assets in the defendant's hands and, before the claim is brought, the defendant successfully invests or exchanges the plaintiff's asset for other more valuable assets. For example, the defendant might sell the plaintiff's shares and invest the traceable proceeds in a painting which turns out to be a masterpiece. The plaintiff's ownership of the shares (whether at law or in equity) allegedly justifies the plaintiff's ownership of the painting. This result effectively compels the defendant to disgorge the profits of the successful investment. This form of disgorgement appears unjustified unless the defendant happens to be a fiduciary or, alternatively, on the arguments advanced later, the defendant is in breach of an obligation of good faith.[73]

To date, and bar one case,[74] the argument in favour of *general* disgorgement of investment gains remains academic.[75] In all the frequently cited cases the defendant is either a fiduciary (so quite different rules apply) or the defendant's investments have not enhanced the value of the plaintiff's assets (so there is no investment gain to be disgorged). The proponents of this form of profits disgorgement justify the remedial response either as 'restitution for wrongs' or as a 'second measure'[76] or 'value surviving' response to the original subtractive unjust enrichment.

The first route is simply not supported by the cases. In the illustration given, conversion of the plaintiff's shares is clearly a wrong, and the defendant's ownership of the masterpiece may well be causally related to that wrong. Nevertheless, the cases on proprietary torts – discussed earlier – confirm that full disgorgement of all traceable gains is not the orthodox remedy.[77] A desire to strip wrongdoers of ill-gotten gains is not enough.

The second route, based on a 'second measure' or 'value surviving' response to the original subtractive unjust enrichment, is equally suspect. It rests on various assumptions or assertions which cannot be defended. Sometimes it is alleged that the plaintiff's shares are being traced into the painting, and that ownership of the first results in ownership of the last. But this ignores the truth about tracing: it is simply a technique for locating the current value of the plaintiff's original asset; it tells us nothing of the plaintiff's claim to the locus. The claim at the end of the tracing exercise remains just the same claim as the one which existed at the

73 Stated in this form, the phrase appears unacceptably expansive; in the later discussion it is in fact quite narrowly defined.

74 *Trustee of the Property of FC Jones & Sons* v *Jones* [1996] 3 WLR 703

75 And general disgorgement is usually supported: see, eg, Birks ch III, but see 352–355; R. Chambers, *Resulting Trusts* (Oxford: Clarendon Press, 1997) 21–27; L.D. Smith, *The Law of Tracing* (Oxford: Clarendon Press, 1997) 156–159; Burrows 59; A. Tettenborn, *Law of Restitution in England and Ireland* (London: Cavendish Publishing, 2nd ed, 1996) 42–43.

76 See Birks ch XI.

77 Unless the defendant is a fiduciary, where an account of profits or constructive trust will often achieve this end. Note that claims to the proceeds of sale of the primary asset do not raise the same issue. If, for example, the shares were sold by the defendant for more than their apparent market value, the plaintiff would certainly be able to recover this sum from the defendant, either as compensatory damages for the tort of conversion (because the court is persuaded that the actual sale price reflects the true market price) or as restitution for unjust enrichment (with the sale price serving to quantify the defendant's incontrovertible benefit)

© The Modern Law Review Limited 1999

beginning of the exercise.[78] Even conceding the (often questionable) existence of a proprietary claim to the original shares, the plaintiff's unjust enrichment claim at the start is only a claim to the shares *in specie* or to their *unjust enrichment value* in the defendant's hands. It remains this at the end of the tracing exercise. In truth, it is the proprietary *nature* of a claim that can persist in traceable proceeds; ownership cannot be assumed. It follows that a plaintiff with an initial proprietary claim to the shares could undoubtedly assert a *lien* over the painting for the value of the shares plus interest. But a further claim to the painting itself (or its entire value) must be justified.[79] Nor will the analogy with purchase money resulting trusts suffice.[80] The intention to provide the funds *as purchase funds* is crucial to the analysis in those cases;[81] only when the defendant is a fiduciary is absence of that intention irrelevant.[82] Nor will it do to say that unauthorised use of the plaintiff's property automatically entitles the plaintiff to the proceeds of that use. That is mere assertion, and depends on the truth of the claim that such gains are *ipso facto* unjust enrichments derived at the plaintiff's expense. As discussed earlier, even the cases claimed as illustrating disgorgement for wrongful use of the plaintiff's assets do not strip the defendant of all derived gains,[83] but only of a 'use value' for unauthorised use of the plaintiff's asset. In short, neither by recourse to second measure claims for restitution nor by disgorgement arguments is it possible for the plaintiff to lay claim to the defendant's successful investment. Of course, the result would be quite different if the defendant were a fiduciary, but then the claim would be based on a different cause of action, relying on a distinct and different breach of duty.

In summary, the view advanced here is that a plaintiff cannot claim the benefits of the defendant's successful investments *simply* because they are derived from assets which were initially the plaintiff's. Unauthorised use (including use to make alternative investments) will justify a restitutionary remedy (including 'use value'), but full disgorgement is, it seems, only justified when such unauthorised use can be additionally characterised as a breach of more onerous and demanding equitable obligations.

78 cf Sir Peter Millett, 'Restitution and Constructive Trusts' (1998) 114 LQR 399, 408: 'The effect of a successful tracing exercise is to confer on the parties the same rights and obligations *mutatis mutandis* in respect of the substituted asset as they ... had to the original asset'. The inference (especially given the reference to *FC Jones & Sons* v *Jones* [1997] Ch 159 immediately following this sentence) is that an initial right to return of the original asset or its value will translate to a right to return of the substituted asset or its value. Although this is often true (eg with claims against fiduciaries), it seems inaccurate as a generalisation. It is important to distinguish between tracing for the purpose of asserting the original restitutionary claim (since identification of substituted value may enable the plaintiff to *secure* the original restitutionary claim against the substituted asset) and tracing to locate the profits of substitutions (which is only relevant if the plaintiff can assert a claim to those profits). Neither the availability of proprietary restitutionary claims, nor the availability of claims to profits (disgorgement) can be assumed; tracing says nothing about either.

79 This, with respect, was the error in *Trustee of the Property of FC Jones & Sons (a firm)* v *Jones* [1996] 3 WLR 703.

80 But see the assertion to the contrary in Chambers, n 75 above, eg at 21, where it is suggested that use of resulting trust principles requires only a slight extension of orthodoxy.

81 As recognised (but not conceded) in Chambers, *ibid* 27, citing, eg, *Aveling* v *Knipe* (1815) 19 Ves 441; *Hussey* v *Palmer* [1972] 1 WLR 1286, 1293; *Calverley* v *Green* (1984) 155 CLR 242, 246; W.J. Mowbray, *Lewin on Trusts* (London: Sweet & Maxwell, 16th ed, 1964) 130; D.W.M. Waters, *Law of Trusts in Canada* (Toronto: Carswell, 2nd ed, 1984) 305; K. Gray, *Elements of Land Law* (London: Butterworths, 2nd ed, 1993) 390; H.A.J. Ford and W.A Lee, *Principles of the Law of Trusts* (Sydney: Butterworths, 3rd ed, 1996) para 2111.

82 eg *Ryall* v *Ryall* (1739) 1 Atk 59, relied on by Chambers n 75 above, eg at 27. Where the defendant is a fiduciary, the resulting (or constructive?) trust is defensible on other grounds. On the arguments advanced later in this article, the same grounds would apply where the defendant was in breach of a duty of good faith.

83 Unless the defendant is a fiduciary, and then, as argued later, the rationale is not restitutionary.

© The Modern Law Review Limited 1999

Unauthorised use of a plaintiff's property: is the restitutionary remedy proprietary?

If the remedy in these cases of unauthorised use of the plaintiff's property is seen as restitution for unjust enrichment rather than as disgorgement for wrongs, does this affect the analysis of whether the remedy is or can ever be proprietary? If the remedy is classified as restitution for wrongs, the disgorgement response is readily regarded as punishing the defendant and as granting a windfall to the plaintiff. In these circumstances it is easily argued that the remedy should not be proprietary,[84] and perhaps even that it should be subordinated to the claims of the defendant's other creditors.[85]

On the other hand, if the remedy is seen as restitution for unjust enrichment, then orthodox restitution scholarship would see the remedy as proprietary if the plaintiff can show an initial proprietary base and can identify an unbroken chain of events linking that base to identifiable substitutions.[86] This will sometimes be possible where the defendant has made unauthorised use of the plaintiff's asset by *exchanging* it for other assets. But this argument is not so easy where the defendant has simply *used* the plaintiff's asset, say by hiring it out. Even if the hiring fees constitute an identifiable fund, careful analysis suggests that the restitutionary remedy will not – and cannot – be classed as proprietary.

What seems to underpin a proprietary restitutionary remedy is the existence of alternative formulations for the initial claim: the defendant is required to return the plaintiff's asset *in specie or* its *unjust enrichment value*. If the asset is no longer in the defendant's hands, but its traceable proceeds can be identified, then the alternative claim can legitimately be secured against those traceable proceeds.[87] But if the plaintiff's unjust enrichment claim is simply for the defendant to pay over the 'use value' of an asset, there is no alternative proprietary formulation: the 'use value' measures enrichment; it does not identify a discrete fund required to be paid over *in specie*.[88] This makes it irrelevant that the use has generated identifiable proceeds; the remedy cannot be proprietary.[89]

Unauthorised use of a plaintiff's property: future developments?

In all of the preceding discussion the focus has been on the use of land or chattels; the use of money has not been considered. It ought to be possible to adopt exactly the same approach if the defendant has made unauthorised use of the plaintiff's money. Where it can be established that a defendant has used money which is the plaintiff's (either at law or in equity),[90] it ought to follow from the arguments advanced here that profits disgorgement would not be required, but that restitution

84 See, eg, R. Goode, 'Property and Unjust Enrichment' in A. Burrows (ed), *Essays on the Law of Restitution* (Oxford: Clarendon Press, 1991) ch 9; R. Goode, 'The Recovery of a Director's Improper Gains: Proprietary Remedies for the Infringement of Non-Proprietary Rights' in E. McKendrick (ed), *Commercial Aspects of Trusts and Fiduciary Obligations* (Oxford: Clarendon Press, 1992) ch 7.

85 See especially P. Jaffey, 'Restitutionary Damages and Disgorgement' [1995] RLR 30, 44.

86 Birks 378–385.

87 See S. Worthington, *Proprietary Interests in Commercial Transactions* (Oxford: Clarendon Press, 1996) ch 8.

88 So the arguments from *A-G for Hong Kong* v *Reid* [1994] 1 AC 324 will not assist.

89 Were this not the case, many torts would provide plaintiffs with secured damages claims eg damages for defamation might be secured against the profits derived by the defendant from the tort. This is not the law.

90 And this will not necessarily be easy: see *Westdeutsche Landesbank Girozentrale* v *Islington London BC* [1996] AC 669, especially the judgment of Lord Browne-Wilkinson.

© The Modern Law Review Limited 1999

of the 'use value' of the money would. The defendant's unjust enrichment is the benefit obtained at the plaintiff's expense from the unauthorised use of the plaintiff's funds. Prima facie, this should be calculated as interest on the original sum: this is the 'use value' of money.[91] If such an approach were adopted, then some of the current difficulties with awards of interest might disappear.[92]

How much further can the analogy can be stretched? Will it accommodate the defendant's unauthorised use of intangible property? Will it stretch to the use (or abuse) of personal rights? Certainly it seems appropriate for the same reasoning to be applied to intangibles, notwithstanding that sometimes there will be additional difficulties in valuing the use-benefit received.[93] It also follows that any developments in the law of personal property which conceded 'property' status to rights previously regarded as purely personal would, in addition, enlarge the scope for application of subtractive unjust enrichment principles. The arena where this seems likely to happen first is with information rights. Information is not regarded as property, yet the push of the information age demands some recognition of the many striking parallels. Such a development would liberate the protection of information from the confines of breach of confidence requirements (where it is the relationship of confidence, not the information *per se*, which is all important) and from the limitations of statutory regimes (where protection is extended only to tightly defined categories). Unjust enrichment principles could then provide an alternative and additional avenue of protection in regulating the unauthorised use of information.

On the other hand, it does not seem appropriate to try to adopt the same reasoning with use (or, more appropriately, abuse) of personal rights.[94] Such rights are designed to achieve more limited ends (avoidance of harm to others; legitimation of expectations), and the remedies are justified and quantified so as to further those ends. All these rights are defined and delimited by the remedies they attract. We may wish for different rights – and other jurisdictions attest to the practical possibility – but it would signify a radical change in English law were it to be suggested that contractual rights and obligations, for example, were premised not only on meeting legitimate expectations but also on restricting self-interested behaviour.[95] This is notwithstanding that some cynical and profitable infringe-

91 In most situations this will be compound interest; simple interest will rarely reflect 'use value'. However, the defendant's innocence may be relevant in determining the appropriate value, views concerning the issues of incontrovertible benefit and subjective devaluation are obviously relevant.

92 This is a controversial area: see *Westdeutsche Landesbank Girozentrale* v *Islington London BC* [1996] AC 669. Also see the discussion in K. Mason and J.W. Carter, *Restitution Law in Australia* (Sydney: Butterworths, 1995) ch 28.

93 Sometimes this will be easy: dividend payments might be seen as part of the 'use value' of shares.

94 This says nothing about cases underpinned by specific receipts. With these cases the remedy is based on conventional notions of subtractive unjust enrichment, not on any purported valuation of the *use* of the plaintiff's personal rights: see, eg, *Astley* v *Reynolds* (1731) 2 Str 915 (money paid under duress to recover pawned goods); *Kettlewell* v *Refuge Assurance Co* [1909] AC 243 (return of insurance premiums acquired through deceit); and other examples cited in Law Commission Consultation Paper, n 36 above.

95 cf R. Nolan, 'Remedies for Breach of Contract: Specific Enforcement and Restitution' in F. Rose (ed), *Failure of Contracts* (Oxford: Hart Publishing, 1997) ch 3, 43–44, who argues that disgorgement is an appropriate response to breaches of contract because the plaintiff has *exclusive* rights to the performance promised, and the defendant who profits from a breach of contract makes those profits by denying the plaintiff that to which he or she is exclusively entitled. But arguably there is a flaw in the initial premise: the defendant has only promised to satisfy the plaintiff's bargained-for expectations, not to subject himself or herself to obligations of self-denial; the latter would convert all contracting parties into fiduciaries. Otherwise there seems to be room for this type of argument only where the contract is specifically enforceable and the plaintiff thereby acquires equitable ownership of an asset which the breaching defendant uses to make the ill-gotten profits.

© The Modern Law Review Limited 1999

ments of personal rights seem to call into question the merits of this restrictive approach. In the next section it is argued that these concerns are better met through principled developments within the law concerning obligations of good faith and loyalty, rather than through the vehicle of unjust enrichment law.

Unauthorised use of a plaintiff's property: summary

To reiterate, the argument presented in this section is that the law of subtractive unjust enrichment is capable of – and, even more than that, is *the* appropriate restitutionary vehicle for – dealing with a defendant's unauthorised *use* of the plaintiff's property. The reason this approach has not been adopted earlier appears to rest, at least in part, on an unstated assumption that an enrichment is only 'at the plaintiff's expense' if the plaintiff is in a worse financial position after the event. This is unduly restrictive. By analogy with unjust enrichment claims relating to provision of services, the true limitation is better seen as requiring only that the enrichment be derived exclusively *from* the plaintiff (whether from the plaintiff's services, assets, or use of those assets). If this approach were adopted, then the existing common law cases currently seen as illustrating disgorgement for wrongs could be divided into two quite separate classes. The first, the proprietary tort cases, would be seen as simple instances of subtractive unjust enrichment. The remedy would be restitution, restricted to the 'use value' of the asset misused; it would not be disgorgement. The second class, comprising the equitable wrongs, would be cases of true disgorgement, but the remedy, as will be seen later, would be underpinned not by a principle against unjust enrichment but by a principle of good faith and loyalty. There would then be no need either for theories defining which wrongs will support disgorgement remedies or for complementary theories justifying limits on the extent of disgorgement required.

Breaches of the obligations of good faith and loyalty: the 'equitable wrongs'

Outside the category of proprietary torts, the other class of cases described as illustrating disgorgement for wrongs comprises the cases on equitable wrongs.[96] Whatever might be said about the classification of remedial responses to proprietary torts, it is unquestionably the case that disgorgement is the orthodox remedy for breaches of fiduciary obligation and breaches of equitable obligations of confidence. The purpose of this section is to advance the idea that if these cases on equitable obligations are the only true disgorgement cases, and if (as is commonly accepted) the disgorgement response in these cases is seen as essential to the effective protection of the relationship, then this provides a rational basis for segregating these cases. They can be seen as forming a coherent class. Within this class, and only within it, the law provides for a disgorgement remedy. The remedy is justified simply as the most effective means of supporting the underlying obligation.

96 Contrary to the arguments advanced here, some might suggest it also necessary to add a class for 'disgorgement for breach of contract' following *dicta* in *A-G* v *Blake* [1998] 1 All ER 833, 844–846, but see n 3 above and the discussion which follows.

© The Modern Law Review Limited 1999

Equitable wrongs: a disgorgement analysis

In advancing this argument, two issues must be addressed. The first is to recognise the difference between the quantification process adopted in these disgorgement cases and that adopted in assessing 'use value' in the proprietary tort cases. The second is to affirm that disgorgement – true disgorgement – is only available where the defendant has breached an equitable obligation of good faith or loyalty;[97] if a different type of obligation is breached, a different form of remedy is ordered.

On the first issue, quantification of 'disgorgement' is analytically distinct from quantification of 'use value', notwithstanding that the calculation may sometimes produce the same monetary sum. The disgorgement remedy, by its terms, requires payment to the plaintiff of all the defendant's ill-gotten gains. However, gains are only ill-gotten to the extent that they are derived from a breach of the obligation in issue, so quantification may sometimes be difficult.[98] Consider a breach of the equitable duty of confidence. If the information wrongfully used by the defendant was vital to the defendant's whole venture, then full disgorgement of the net profits of the venture is the appropriate disgorgement measure.[99] On the other hand, if the information was not particularly valuable to the venture, being information which might have been obtained for a fee from alternative sources, then disgorgement of the profits of the breach requires careful (and often difficult) assessment of the proportion of the profits realistically attributable to the use of the information.[100] The plaintiff is not necessarily limited to the 'use value' or 'licence fee' for the information, but is prima facie entitled to a proportionate share of the profits.[101] In every case, the aim, in quantifying the disgorgement remedy, is to strip the defendant of *all* ill-gotten gains, but no more and no less.

If the remedy in these equity cases is full disgorgement of the defendant's ill-gotten gains, then the next issue is what justifies such a response? A common understanding of this area of the law is that the disgorgement response is directed at protecting particular types of relationships.[102] The law recognises certain relationships as warranting legal protection and provides that protection by imposing additional obligations of good faith and loyalty on one of the parties. These additional obligations arise by operation of law, not by agreement between the parties. In this sense they parallel obligations imposed by the law of tort or unjust enrichment, and differ from obligations imposed by contract. But what sets these obligations apart from contract, tort and unjust enrichment is that their focus is on the social value of the relationship, not on the moral obloquy of the defendant or the need to protect the plaintiff from harm. A defendant may be found in breach of these equitable obligations notwithstanding that the breach was committed in

97 The difficult issue of whether this disgorgement remedy ought to be proprietary is not considered

98 See nn 40–42 above.

99 As in *Peter Pan Manufacturing Corp* v *Corsets Silhouette Ltd* [1963] 3 All ER 402, where, notably, the remedy was for breach of the equitable obligation of confidence, not the contractual one. Also see *Lac Minerals Ltd* v *International Corona Resources Ltd* [1989] 61 DLR (4th) 14.

100 See, eg, the approach adopted in *Celanese International Corp* v *BP Chemicals Ltd, The Times*, 5 November 1998 (Mr Justice Laddie), a patent infringement case. Also see *Attorney-General* v *Guardian Newspapers (No 2)* [1990] 1 AC 109, where *The Sunday Times* was not required to disgorge the entire profits of the relevant day's publication, but only the profits which could reasonably be attributed to the wrong.

101 See *Celanese International Corp* v *BP Chemicals Ltd, The Times*, 5 November 1998. Of course, sometimes a 'licence fee' *is* the appropriate measure of the profit the defendant has gleaned from the breach: see *Seager* v *Copydex (No 2)* [1969] 1 WLR 809, 813 *per* Lord Denning MR.

102 The notion has a long history: see, eg, *Keech* v *Sandford* (1726) Sel Cas Ch 61; *ex parte James* (1803) 8 Ves 337; *Bray* v *Ford* [1896] AC 44.

© The Modern Law Review Limited 1999

ignorance,[103] or in good faith, or that it did no harm to the plaintiff, and may even have provided a benefit.[104] It is clearly the relationship which is valued, not any underlying property or information, because the obligations may be imposed in the absence of any property[105] and regardless of the economic value of the information.[106] The fact that it is equity's social concern to protect certain relationships is further reinforced by the availability of defences which appeal to even higher social interests; for example, obligations of confidence can be breached where to do so is in the public interest.[107]

These protected relationships are seen as sufficiently important that the remedy is designed, as far as remedies can be, to ensure that the imposed obligation is *not* breached, not that a breach does no harm. The aim is to exact particular standards of conduct in protected relationships; to this end, the relevant law is concerned with proscribing certain activities, not with precluding particular outcomes. The appropriate remedial response for breaches of these equitable obligations is disgorgement because this is the remedy which best supports the legal obligation being enforced. Outside the arena of these equitable obligations, other remedies will suffice to support the ends which are desired.[108] Across the legal system, this differential approach is efficient: it imposes the minimum degree of legal intervention necessary to achieve the ends desired; this is what justifies the imposition of different remedies for breach of different obligations, and why it seems incongruous to suggest that the entire remedial menu of the common law should be available to remedy any cause of action.[109]

---

103 Ignorance that the *act* is in breach of the defendant's equitable obligation is irrelevant: see, eg, *Seager* v *Copydex* [1967] 1 WLR 923 However, it seems that the defendant must know (or must have once known) that the *relationship* itself exists. Whether this requires actual knowledge or constructive notice (as is commonly the case in applying equitable doctrines) is a difficult point: see eg, *Coco* v *AN Clark (Engineers) Ltd* [1969] RPC 41, 47–48 *per* Megarry J, *AG* v *Guardian Newspapers Ltd (No 2)* [1990] 1 AC 109, 281 *per* Lord Goff (*obiter*); Goff & Jones 685–686. This knowledge requirement ensures that resulting and constructive trustees may sometimes escape liability to disgorge: see *Westdeutsche Landesbank Girozentrale* v *Islington London BC* [1996] AC 669, [1996] 2 WLR 803, 828–829 *per* Lord Browne-Wilkinson.

104 The most familiar examples are *Boardman* v *Phipps* [1967] 2 AC 46; and *Regal (Hastings) Ltd* v *Gulliver* [1942] 2 All ER 378.

105 eg the fiduciary obligation between solicitor and client exists independently of any property which might be subjected to the relationship.

106 eg *Prince Albert* v *Strange* (1849) 2 DeGex & Sm 652, 64 ER 293; *Argyll* v *Argyll* [1967] 1 Ch 302, 332 *per* Ungoed-Thomas J. In *Seager* v *Copydex Ltd* [1967] 1 WLR 923, 931 *per* Lord Denning, the jurisdiction was seen as founded on broad principles of equity which required the defendant not to take unfair advantage of the plaintiff, and not on some property basis; also see *Schering Chemicals Ltd* v *Falkman Ltd* [1982] QB 1, 28 *per* Shaw LJ, and Goff & Jones 684 and the cases cited there. This view that the obligation is imposed to effect a social purpose is reinforced by the special considerations which apply where the relationship is between government and subject: the subject is only bound in confidence if disclosure would *damage* the public interest; see, eg, *AG* v *Jonathan Cape Ltd* [1976] QB 752; *Commonwealth of Australia* v *John Fairfax & Sons Ltd* (1980) 147 CLR 39 (Aust).

107 eg *Lion Laboratories Ltd* v *Evans* [1985] QB 526; *D* v *NSPCC* [1977] 1 All ER 589, 594 *per* Lord Diplock.

108 This is much narrower than the view advanced by I.M. Jackman, 'Restitution for Wrongs' [1989] CLJ 302: in defining which wrongs will give rise to 'restitution for wrongs' and which will not (since he concedes that not all profitable wrongdoing merits this response), Jackman concludes that the remedy is available to protect a variety of 'facilitative institutions', namely private property, relationships of trust and confidence, and (with some qualification) contracts. Under each of these heads, Jackman gives illustrations of cases where the court has awarded 'restitution for wrongs' (although there is no attempt to differentiate between full disgorgement and 'use value'). His approach is inductive, but it does not demonstrate why only (or all) these institutions *need* this form of protection to survive. Property and contract appear adequately protected by alternative means.

109 Although see *Aquaculture Corp* v *New Zealand Green Mussel Co Ltd* [1990] 3 NZLR 299, 301.

© The Modern Law Review Limited 1999

Equitable wrongs: future developments?

If these equitable obligations – and no others – give rise to disgorgement remedies, then the fear is that courts will be tempted to re-characterise torts and breaches of contract as breaches of fiduciary (or other equitable) obligation. The fear is not fanciful. Several Canadian decisions, in particular, illustrate the difficulties which can arise with an instrumental use of the fiduciary label.[110] That approach is not part of the thesis advanced here. In fact, if the proper function of these equitable obligations were fully appreciated, then their roles might be seen as confined to strictly limited circumstances: 'fiduciaries' and 'confidants' are only needed where the law sees the plaintiff as entitled to the protection that such a role will deliver;[111] that ought to be so only when the relationship cannot function effectively without such additional safeguards. To define these circumstances restrictively should be no more difficult than defining obligations of care restrictively, and just as necessary.

Nevertheless, an approach based on obligations of good faith and loyalty does provide scope for principled developments in the law. The imposition of these equitable obligations is sometimes seen as a sign of the increasing sophistication of a legal regime.[112] That development continues. There are already signs of a desire to import an additional 'good faith' obligation into some, if not all, dealings between individuals.[113] The standard of conduct being mooted is not one which demands self-denial (ie a fiduciary standard) in all relationships; that would be inappropriate. It probably does not even go so far as to require positive regard for the legitimate interests of others (ie a conventional good faith standard). What it seems to demand is avoidance of activities which display cynical *disregard* for the legitimate interests of others – true, this is not as demanding as a strict obligation of good faith;[114] it is perhaps more correctly styled as an obligation not to act with contumelious bad faith.[115]

If the law were to develop to recognise that all (or perhaps only some) relationships warrant the additional protection which might be derived from an obligation of good faith – or at least an obligation not to act with contumelious bad faith – then this obligation would be added to the stable of existing equitable

110 See, eg, *Norberg* v *Wynrib* (1992) 92 DLR (4th) 449 (Can) (fiduciary relationship between doctor and patient); *M(K)* v *M(H)* (1993) 96 DLR (4th) 449 (Can) (fiduciary relationship between child-abuser and victim).

111 P.D. Finn, 'Fiduciary Law and the Modern Commercial World' in E. McKendrick (ed), n 84 above, 9; and P.D. Finn, 'The Fiduciary Principle' in T.G. Youdan (ed), *Equity, Fiduciaries and Trusts* (Toronto: Carswell, 1989) 1, 4.

112 See eg, J.P. Dawson, *Unjust Enrichment: A Comparative Analysis* (Boston: Little Brown, 1951) 39–40

113 See, eg, P.D. Finn, 'The Fiduciary Principle' n 111 above, especially 10–24; F. Kessler and E. Fine, 'Culpa in Contrahendo, Bargaining in Good Faith, and Freedom of Contract: A Comparative Study' (1964) *77 Harvard Law Rev* 401; B.J. Reiter, 'Good Faith in Contracts' (1983) 17 *Valparaiso UL Rev* 705; H.K. Lucke, 'Good Faith and Contractual Performance' in P.D. Finn (ed), *Essays on Contract* (Sydney: Law Book Co, 1987) ch 5.

114 Even in civilian jurisdictions, where such an obligation is common, it is attenuated to meet particular circumstances. It can, however, justify disgorgement for certain bad faith breaches of contract: see, eg, *Adras Building Material Ltd* v *Harlow & Jones Gmbh* [1995] RLR 235 (Israel), noted by Nolan, n 95 above, 58, n 83.

115 Awards of exemplary damages in similar circumstances demonstrate that the issue is already a concern of the law, although so far it is only dealt with in a rather unstructured and *ad hoc* manner: see *Rookes* v *Barnard* [1964] AC 1129 for the jurisdictional limitations in England. In other Commonwealth jurisdictions, the remedy is available for torts which display the defendant's wanton and contumelious disregard of the plaintiff's rights: see A. Burrows, 'Reforming Exemplary Damages' in Birks (ed), n 4 above, 165 and references cited there.

© The Modern Law Review Limited 1999

obligations and enforced by imposition of a disgorgement remedy. The 'equitable' tag is used simply to underline the fact that the obligation is directed at protecting relationships; it is not, for example, a tort obligation directed at protecting affected individuals from harm (and compensating that harm). It also serves to underline the fact that the disgorgement remedy is not simply parasitic: disgorgement does not follow from proof that a tort, for example, has been committed; it would only follow from proof that an obligation of good faith had been breached (by acting with cynical disregard for the plaintiff's rights).

Conclusions

The thesis advanced in this article can be simply stated: it is that there is no dependent or parasitic category of restitution for wrongs, no category which imposes a disgorgement remedy on wrongdoing defendants as an alternative to the 'normal' remedy for the wrong. Broadly speaking, rights or obligations do appear to map onto discrete remedies. Torts usually give rise to compensatory damages;[116] breaches of contract to expectation damages; unjust enrichments to restitution. The only obligations which seem to give rise to a disgorgement remedy are equitable obligations of good faith and loyalty. This is not to deny that the same set of facts can give rise to alternative claims: a plaintiff may be in a position to advance claims in contract, tort, unjust enrichment and obligations of good faith and loyalty. What the thesis does deny is that a claim in tort, for example, could result in *either* compensatory damages or disgorgement.[117] In order to establish an entitlement to disgorgement, it seems the plaintiff must do more than show that a tort (even a tort of a restricted class) has been committed; the plaintiff must show that the defendant has also breached an equitable obligation of good faith or loyalty.

This follows from the observation that cases which are commonly assigned to the category of restitution for wrongs might usefully be reassigned; they do not seem to fit easily within their current classification. Some cases – the 'proprietary torts' – are better seen as cases where the plaintiff is entitled to restitution for subtractive unjust enrichment for unauthorised use of the plaintiff's property. The measure of the defendant's unjust enrichment is the 'use value' of the asset to the defendant. This claim in unjust enrichment is an alternative to the plaintiff's claim for compensatory damages for the tort, but it is not dependent on commission of the tort; it is an independent claim requiring only proof of the defendant's unauthorised use of the plaintiff's property (ie use without the plaintiff's consent). Other cases – the 'equitable wrongs' – are true examples of disgorgement, but they are better seen as cases falling entirely outside the law of unjust enrichment. The reversal of unjust enrichment (even unjust enrichment from wrongdoing) is not their rationale. The obligations in issue are not like obligations in tort or contract or unjust enrichment; these obligations merit a class of their own to recognise this difference.

116 Sometimes punitive damages may be awarded. It seems desirable to retain this remedial option even if the thesis advanced here is accepted. It is true that many cases where punitive damages are currently awarded could be re-analysed as cases where an obligation of good faith has been breached and disgorgement is therefore appropriate; however, there will still be cases where the defendant has not profited (or not greatly), and yet punitive damages seem warranted (eg deliberate pollution, or deliberate defamation). Significantly, the objective underpinning the award of punitive damages, even in these cases, remains protection of socially desirable relationships.

117 cf Birks 316.

© The Modern Law Review Limited 1999

This suggested reclassification is grounded in an attempt to rationalise the principles governing disgorgement. It recognises two significant issues. First, it identifies the importance of the obligation to make restitution of *all* unjust enrichments. Use of the plaintiff's property is a source of enrichment just as much as use of the plaintiff's services. An unjust enrichment is not derived solely from possession of the property itself. Secondly, it recognises that the remedy of disgorgement is mediated by different legal imperatives from those governing restitution, or compensatory or expectation damages. This can only be recognised – and should be recognised – by dissociating these cases from cases in the other categories. For good reason the common law responds differently to torts, unjust enrichments, breaches of contract and breaches of equitable obligations of good faith and loyalty.

© The Modern Law Review Limited 1999

LEGISLATION

Unfinished Business: The Scotland Act 1998

*Noreen Burrows**

The Scotland Act 1998, which received the Royal Assent on 19 November 1998, is one of a series of Acts whose aim is to modernise the United Kingdom (UK) constitution.[1] In Scotland, Northern Ireland and Wales, regional parliaments or assemblies are to be established with the goal of decentralising power. Devolution of power to the 'regions' of the UK is not symmetrical. The Scottish settlement differs in fundamental respects from the Northern Irish or Welsh settlements and they in turn differ. Asymmetry in our constitutional order should not be seen as a weakness in itself.[2] In many ways the constitutional settlements reflect the aspirations and needs of the people in the different regions of the UK. Aspirations and needs change. John Smith called the devolution project 'unfinished business'. This is how the Scotland Act should be viewed. It is not the final word but it is the first step in the creation of a new constitutional settlement for Scotland and the UK. Many issues remain to be resolved such as the relationship between the devolved governments and parliaments and existing UK and European institutions. These will be resolved by the development of new constitutional or political conventions which in turn will require a fundamental review of our constitutional theory. Devolution is not a local matter. It has implications for the entire UK constitutional settlement.

This article examines the main provisions of the Scotland Act in the context of the debates taking place in Scotland about devolution and the creation of the new Scottish Parliament. The Scottish Parliament is the visible symbol of constitutional change in Scotland and it will be the focus for change in the future.[3] The new Parliament is not a resumption of the old Scottish Parliament. It is a creation of the UK Parliament and as such is subordinate to it. The old Scottish Parliament was the parliament of an independent nation; the new Parliament is the parliament of a region within a larger nation state. As such it is comparable to the parliaments of the regional autonomies of Spain or the German Lander with the major difference that these parliaments operate in a constitutional framework that recognises the legal personality of the region and guarantees the constitutional protection of the regional governments. In the UK new regional governments and institutions do not

* School of Law, University of Glasgow.

I am grateful to Tony Prosser for his comments on an earlier draft of this article.

1 *The Herald* newspaper kept some statistics on the passage of the Act. It reported that 'it took Parliament 191 hours and 43 minutes of debate during 32 days at Westminster, 15 in the House of Commons and 17 in the House of Lords. More than one million words were spoken and 2025 amendments were tabled, 670 from the Government'. *The Herald* 18 November 1998.

2 M. Keating, 'What's wrong with asymmetrical government?' in H. Elcock and M. Keating (eds), *Remaking the Union. Devolution and British Politics in the 1990s* (London: Frank Cass, 1998).

3 For the last thirty years the focus of constitutional debate in Scotland has been on the creation of and need for a Parliament. For a collection of materials on this debate see L. Paterson, *A Diverse Assembly: The Debate on a Scottish Parliament* (Edinburgh: EUP, 1998).

© The Modern Law Review Limited 1999 (MLR 62 2, March). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

have this constitutional protection as yet but if they are successful in operating within a devolved framework it is at least arguable that the logic of devolution requires the creation of a constitutional order which guarantees their autonomy.

Background

Both the Scottish and English parliaments were abolished by the Acts of Union of 1707. A new parliament, the Parliament of Great Britain, was established merging the two parliaments. The Acts of Union served to protect some aspects of Scottish life and culture, most notably its legal system and the Church of Scotland as the established church in Scotland.[4] However it has often been repeated that Scotland is one of the few independent legal orders lacking a legislature. Scottish legislation, in time, came to have its own procedures in the Westminster Parliament but these were inadequate mechanisms for securing popular support and accountability for government policy in Scotland. At the same time, the powers of the Scottish Office were gradually increased, so much so that complaints were levelled that the Scottish Office was free to act like the colonial governing power of some far-flung colony. The Scottish Office was unaccountable to the Scottish people and ignored by Westminster. In Scotland, there was a real democratic deficit.

The Royal Commission on the Constitution (1969–1973) had recommended that legislative competence be devolved to Scotland.[5] The Labour Government enacted the Scotland Act 1978. That Act was to devolve certain specified areas to a Scottish Parliament, which was to be elected following the result of a referendum, provided that the referendum achieved a positive vote of 40 per cent of the electorate in favour of devolution. The referendum target was not achieved with only 32.5 per cent of the electorate voting in favour of devolution. The 1978 Scotland Act was, therefore, repealed. In May 1979 the election of a Conservative government opposed to devolution meant that devolution was no longer on the official political agenda.

Within Scotland, however, devolution remained a live issue particularly in the light of dwindling political support for the Conservatives culminating in the May 1997 election when no Conservatives were returned from Scottish seats.[6] The Campaign for a Scottish Parliament continued its work and established a Committee, which, in July 1988, recommended the formation of a Scottish Constitutional Convention 'to make plans for the future governance of Scotland'.[7] In early 1989, a cross party group was established to discuss ways of moving the project forward. The Conservatives refused to participate and the Scottish Nationalist

4 For an analysis of the Acts of Union and a summary of some of the literature on them see Michael F. Addo and Veronica M. Smith, 'The relevance of historical fact to certain arguments relating to the legal significance of the Acts of Union' (1998) *Juridical Review* 37.

5 *Report of the Royal Commission on the Constitution 1969–73*, Cmnd 5460 (1973). For a succinct history of attempts to change Scottish constitutional arrangements prior to the establishment of the Royal Commission and more generally on Scottish politics see A. Brown, D. McCrone and L. Paterson, *Politics and Society in Scotland* (Basingstoke: Macmillan, 2nd ed, 1998).

6 Paterson identifies four themes as the dominant concerns of the debate in Scotland: good and effective government, nationalism, the importance of community and civil society in Scotland, and negotiated autonomy. See Paterson in n 3 above

7 The background to the work of the Scottish Constitutional Convention can be found in *Scotland's Parliament: Scotland's Right*, the report presented to the Scottish people by the Convention (Edinburgh, 1995). Information about the Scottish Constitutional Convention can be found on its web site http://www.almac.co.uk/business-park/scc/backg.htm. The text of the Claim of Right is in the 1995 report and on the web site.

 © The Modern Law Review Limited 1999

Party (SNP) withdrew support preferring to continue its fight for independence. At its inaugural meeting in March 1989 the Scottish Constitutional Convention adopted a *Claim of Right for Scotland* echoing earlier Claims of Right. The modern version acknowledges 'the sovereign right of the Scottish people to determine the form of government best suited to their needs'. It states that the purpose of the Scottish Constitutional Convention is 'to agree a scheme for an Assembly or Parliament for Scotland', to encourage popular support for it and to 'assert the right of the Scottish people to secure the implementation of the scheme'.

Membership of the Scottish Constitutional Convention was drawn from the political parties (Conservatives and SNP excepted although individuals within these parties did support the movement), representatives of local authorities, the churches, civic organisations including women's groups, trade unions and business organisations. It claimed to be 'the most broadly representative organisation in Scotland'.[8] The Scottish Constitutional Convention met in plenary sessions and in working groups until November 1990 when it presented its report *Towards Scotland's Parliament*. At that stage, there was not agreement on the method of election or on certain other matters and an independent commission was established under the convenership of Joyce McMillan to report on these matters. The findings of the McMillan report were incorporated into the conclusions of the 1990 report and a final document, *Scotland's Parliament; Scotland's Right* was presented to the Scottish people on St Andrew's Day 1995.

The involvement of the Labour Party throughout this process means that the 1995 report has a particular significance in relation to the Labour Party's proposals for devolution. The Labour Party was elected on the basis of a pledge to introduce a devolved parliament for Scotland and has made frequent references to the work of the Scottish Constitutional Convention and the principles underlying its work. An understanding of *Scotland's Parliament: Scotland's Right* is therefore crucial to an understanding of the Scotland Act despite the fact that there are some significant differences between them. In July 1997, the new government issued its White Paper, *Scotland's Parliament*, which set out to translate the Labour Party's commitment 'into a sound and durable constitutional settlement'.[9] Whilst acknowledging a debt to the Scottish Constitutional Convention, the white paper reorients the debate on devolution. Whereas the Scottish Constitutional Convention and the Claim of Right had sought to anchor a new political settlement in a concept of sovereignty based on the will of the people, the White Paper states unequivocally that:

> the UK Parliament is and will remain sovereign in all matters: but as part of the government's resolve to modernise the British constitution Westminster will be choosing to exercise that sovereignty by devolving legislative responsibilities to a Scottish Parliament without in any way diminishing its own powers.

The consequences of this approach are discussed below in considering the provisions of the Scotland Act. At this stage, it is perhaps sufficient to note that, in Scotland the 'doctrinal incubus of parliamentary sovereignty' has not gone uncontested.[10]

Before introducing the Scotland Bill, the government decided to ascertain the wishes of the Scottish electorate on devolution. The reasons for holding a

8 Although its membership was overwhelmingly male, middle class and white. See Brown *et al* n 5 above.

9 *Scotland's Parliament* Cm 3658 (1997).

10 See Brown *et al* n 5 above, 46. See also, for example, the work of Neil MacCormick on sovereignties in conflict in *Scottish Devolution: Some Comments*, paper presented to the Society of Public Teachers of Law seminar, 18 October 1997 and the literature reviewed in Addo and Smith, n 4 above.

© The Modern Law Review Limited 1999

referendum were outlined by the Secretary of State, Donald Dewar, in the debates on the Referendums (Scotland and Wales) Bill 1997. He argued that:

> There is advantage in popular consent. In our unwritten constitution, popular consent gives a certain legitimacy. In a sense it is a way to build the devolution scheme into the system and to give it roots ... If we get the right result, we have moral authority to speed the passage of devolution.[11]

Donald Dewar had, of course, been a signatory to the Claim of Right and therefore had some attachment to the notion of the 'settled will of the people'. His claim that popular consent assists in providing legitimacy reflects the Scottish debate and perhaps suggests a slight ambivalence towards any theory of the sovereignty of the UK Parliament in political if not legal terms.

The Referendums (Scotland and Wales) Act 1997 provided that a referendum was to be held on 11 September 1997. In Scotland, two questions were put to the electorate; the first asked whether there should be a Scottish parliament and the second whether that parliament should have tax varying powers. The Labour government rejected SNP demands for a third question on independence. It also rejected Conservative demands that the timing of the referendum should be delayed until after the debate on the contents of the Scotland Bill had taken place. In 1978, the referendum had taken place after the Scotland Act had received Royal Assent. The government's response was to state that if there were not significant support for devolution then it would be pointless to waste valuable parliamentary time. The results of the referendum reflected widespread popular support for the creation of a parliament and for that parliament to have tax varying powers. On a turnout of 60.4 per cent, 74.3 per cent voted in favour of the creation of a Scottish parliament with 25.7 per cent against and 65.6 per cent voted in favour of tax varying powers with 36.5 per cent against. From this time on, the Labour government was able to quote the settled will of the people in support of its devolution proposals. Accordingly the Scotland Bill was published in December 1997.

The Scottish Parliament

Section 1(1) of the Scotland Act provides that 'there shall be a Scottish Parliament'. Thereafter, sections 1(2) to 43 elaborate on its creation, its powers and its functions. The Scottish Parliament will perform several tasks; it may debate any matter national or international, it will have a limited legislative competence, it will provide the Scottish Executive by electing a First Minister and thereafter approving his/her choice of Ministers and law officers, it will call the Scottish Executive to account, it will have limited tax varying powers and the power to approve a Scottish budget.

11 HC Deb vol 294, col 720 21 May 1997 The Scottish Office has a very informative web site where all the debates on the Bill, background papers, press releases and details of the referendum can be found. See http://www.scottish-devolution.org.uk/frame/htm. Space precludes a discussion of the constitutional implications of the use of referendums by the present government The White Paper on the Scottish Parliament attaches great importance to the unchanging nature of the sovereignty of parliament. However it might be difficult politically for any future government to attempt to alter significantly the institutional framework established by the Scotland Act without ascertaining the will of the people.

© The Modern Law Review Limited 1999

Members of the Scottish Parliament

The Act provides for two types of Members of the Scottish Parliament (hereafter MSPs); constituency MSPs and regional MSPs. One constituency MSP is to be elected per constituency. The Act defines constituencies as being the current Westminster parliamentary constituencies except for Orkney and Shetland which are to be separate constituencies. For the purposes of the first elections for the Scottish Parliament, therefore, there will 73 constituencies (Schedule 1). The 73 constituency MSPs will each be elected on the basis of a simple majority vote (first past the post). In addition, there are 56 MSPs elected by an additional member system from 'regions' of Scotland. These regions are defined in terms of the old European constituencies as defined in the European Parliamentary Constituencies (Scotland) Order 1996.[12] There are therefore eight such regional constituencies each returning seven regional MSPs. This makes a total of 129 MSPs with constituency MSPs being in the majority.

Schedule 1 ties the number of constituencies for the Scottish Parliament to the Westminster constituencies. These provisions are not uncontroversial. Section 86, which relates to arrangements at Westminster, amends the Parliamentary Constituencies Act 1986. The effect of the amendment is to remove the guaranteed minimum number of Scottish seats in the House of Commons. That minimum had been 71 seats. Section 86 also directs the Boundary Commission for Scotland, in calculating the number of seats for Scotland in Westminster, to apply the same electoral quota as is used in England. This will reduce the number of Scottish seats in the House of Commons by some 15–20 seats. In itself this is not controversial. The White Paper had stated that such a review would be necessary. It is a response to the so-called West Lothian question raised initially by Tam Dalyell about the inequity of Scottish MPs being able to discuss, and possibly determine, law and policy which would apply solely to England and Wales, such inequity being compounded by a numerical over-representation of Scotland in the House of Commons.[13] What is problematic is the *linkage* between the Westminster and Scottish parliamentary constituencies. If the number of members of the House of Commons decreases so too does the number of MSPs. Furthermore, because the Boundary Commission is required by Schedule 1 to the Scotland Act to attempt to maintain a ratio of 56 regional seats to 73 constituency seats in so far as this is reasonably practicable, the number of regional seats will be similarly reduced perhaps by one or two members per region. This problem would be further compounded if the Jenkins proposals on electoral reform were adopted for the UK. Under Jenkins the number of UK, and hence Scottish, constituencies would be even further reduced.[14] The House of Lords defeated the government twice on this issue but the amendments proposed by the Lords were not accepted.

Apart from their method of election, the Act draws no distinction between regional and constituency MSPs and at this stage it is not clear what their respective roles will be. Much will depend on the outcome of the elections. The idea had been mooted of paying regional MSPs less than constituency MSPs thereby implying a lesser role for regional members. Questions of remuneration

12 The closed list system for election to the Scottish parliament was not opposed in the House of Lords to the extent of blocking the legislation as their Lordships did with the European Elections Bill.

13 Tam Dalyell, *Devolution, The End of Britain?* (London. Jonathan Cape, 1997) 247

14 No similar linkage is made in the scheme for Northern Ireland. The Northern Ireland Assembly will have 108 seats. In the 1920s the number of MPs returned to Westminster was reduced when a devolved government was first established. By 2005, the Northern Ireland Assembly and the Scottish Parliament will be of equivalent size.

© The Modern Law Review Limited 1999

and pensions, of both MSPs and Ministers, however, will fall to be determined by the Parliament itself once it is elected (sections 81 to 83). One possible difference might be in respect of allowances to be made for conducting constituency work.

Methods of election

The Scotland Act provides that each elector has two votes. The first vote is cast for the constituency MSP and the second is for a registered political party that has submitted a list, or for an individual candidate. It is only after the constituency members have been elected that the regional votes come into play (section 7). To determine the allocation of regional seats the total number of votes cast for a party is divided between the number of seats gained in the constituency poll plus one and this process is repeated until all the regional seats are filled. The political parties list their candidates (up to a maximum of 12) and regional seats are allocated to candidates in the order of party preference (sections 7 and 8). Section 5 provides detailed rules as to whether candidates may offer themselves for election on different lists and/or for different constituencies.

The reference in section 5(9) of the Scotland Act to a registered political party presupposes that such a register exists. In the UK there was no legislation on the registration of political parties until the adoption of the Registration of Political Parties Act 1998. That Act creates a register to be held by the Registrar created under the Companies Act 1985. Parties must register annually if they wish to put forward candidates for parliamentary elections, elections to any of the devolved institutions and for European and local government elections. The Act sets out the information which must be provided in order to register and lays down limits as to how political parties might be registered in terms of the name of the party and the use of any emblems.[15]

Qualifications

A candidate need not be Scottish or even resident in Scotland to stand for election. All UK citizens (including peers and priests), Commonwealth citizens, Irish citizens and EU citizens resident in the UK may stand. An individual who is disqualified from being a Member of the House of Commons either by virtue of the House of Commons Disqualification Act or for any other reason is equally disqualified from being an MSP (section 15). Further categories of persons disqualified to stand might be specified by Order in Council. The seats of persons who are, or become, disqualified become vacant. Section 18 gives jurisdiction to the Court of Session to issue a declarator in proceedings in which the qualification of disqualification of an individual is in question.

The question of the dual mandate is not settled by the Scotland Act. The Scottish Constitutional Convention proposed that a dual mandate should not be possible after the first parliament but that for the first parliament it should be possible for MPs, MEPs or councillors in local authorities to continue in post until their next election. The Act is silent on this question except to say that where an individual is also a member of another parliament then that person's salary should be abated (section 82). In practice, it will be difficult for an individual to serve in both parliaments but this is a matter for the individual concerned and his/her political

15 The Registration of Political Parties Act is also intended to remedy the problem raised in *Sanders and another* v *Chichester and another* Queen's Bench Division (Election Court) *The Times* 2 December 1994 where a candidate attempted to stand as a Literal Democrat.

© The Modern Law Review Limited 1999

party. A number of existing MPs have been selected as potential candidates for the Scottish Parliament, including the current Secretary of State and the devolution minister. If these candidates are all elected this might precipitate a rash of by-elections in Scotland for Westminster.

Candidates may be republicans or monarchists but, once elected, they are required to take the oath of allegiance, as are members of the Scottish Executive. The form of the oath is specified in the Act as being the form provided by the Promissory Oaths Act 1868. The Consultative Steering Group (CSG) is recommending that the oath may be taken in English or in Gaelic.[16]

Selection processes

The selection processes used by the political parties have not been without controversy. One of the reasons for this controversy is the absence in the Act of a mechanism to ensure gender parity. The Scottish Constitutional Convention had envisaged a parliament in which there 'should be equal representation of men and women'. Indeed, this was said to be 'one of the key principles' underlying the proposals. The political parties engaged in the work of the Convention entered into an Electoral Agreement by which they agreed 'to select and field an equal number of male and female candidates' in the elections taking into consideration both the constituency and the regional lists.[17] The White Paper seemed to reflect this commitment by stating that the Government 'attach great importance to equal opportunities for all' but merely urged the political parties to have this in mind in the selection processes. The Government indicated early on that it had no intention of legislating to ensure equality in the Scottish Parliament and there is no provision for it in the Act.

Each party has used different selection processes but only the Labour Party has used a semi-formal mechanism of 'zipping and twinning' constituencies to try to ensure that men and women are equally represented.[18] It is unlikely that the first Parliament will have anything like 50:50 representation of men and women. If this is the case then the Constitutional Convention has recommended that the Parliament itself should conduct an enquiry with a view to review and amend the electoral system. Without a firm commitment to equal opportunities on the part of the political parties in Scotland it is unlikely that such a review will take place.

Term

Unlike Westminster, the Scottish Parliament has a fixed term of four years, with elections being held, under normal circumstances, on the first Thursday in May in every fourth year, although the first poll will be held on a day to be determined by the Secretary of State (sections 2–4). He has indicated that the first elections will take place on 6 May 1999 and a detailed Order to this effect will be issued in spring

16 The work of the CSG is discussed below.

17 For details and text of the Electoral Agreement see *Scotland's Parliament: Scotland's Right*, n 7 above. There are a number of web sites which deal with the 50:50 campaign: see the Engender web site http://www.engender.org.uk. Engender, a Scottish women's organisation has published an information pack *Women and the Scottish Parliament* (Edinburgh: Engender, 1997) The positions of the political parties are outlined in this pack.

18 The Labour Party interviewed potential candidates to draw up a list of approved candidates. Constituencies of similar 'winnability' were twinned and invited to select a man and a woman. Constituencies then decided, following an agreed formula, who would stand in each constituency. See Rosemary McKenna in 'Selection Procedures for the Scottish Labour Party Panel for the Scottish Parliamentary Elections' (1998) 6 *Scotland Forum* 3.

© The Modern Law Review Limited 1999

1999. 6 May is also the date for local authority elections in Scotland. It is anticipated that the Parliament will meet for the first time mid-May 1999 and assume full powers on 1 July 1999 with the Queen performing the opening ceremony.

Legislative competence

The Scotland Act 1998 differs from the Scotland Act 1978 in its approach to the definition of devolved matters. The earlier Act had listed devolved areas and reserved the rest for Westminster. This approach had been criticised as being cumbersome and as potentially requiring repeated amendment. The Scotland Act 1998 more sensibly chooses to confer a general power on the Scottish Parliament to make laws to be known as Acts of the Scottish Parliament[19] whilst imposing some limitations on its powers. Section 29 provides that an Act of the Scottish Parliament is not law if it purports to have effect outside Scotland, if it relates to reserved matters, or if it would have the effect of modifying the Acts of Union, the European Communities Act 1972, the Human Rights Act 1998 and most of the provisions of the Scotland Act itself (Schedule 4), or if it is incompatible with any of the rights provided for by the European Convention of Human Rights or with Community law, or if it would remove the Lord Advocate from his position as head of criminal prosecution and investigation of deaths in Scotland.

The reserved matters are set out in Schedule 5. They are divided into general and specific reservations. The general reservations cover the broad headings of the constitution, political parties, foreign affairs, public service, defence and treason. The list of specific reservations is longer and comprises twelve heads including financial and economic matters, home affairs and trade and industry. Within these headings are listed a number of specific areas. Only those aspects of the policy areas listed in the Schedule are reserved. The dividing line between devolved and reserved matters is not absolutely fixed. Section 30(2) provides that Schedule 5 can be modified by an Order in Council where is it necessary or expedient to do so provided that the draft Order has been approved by both the UK and the Scottish Parliaments (Schedule 7). Matters that are currently devolved may, therefore, become reserved but equally possibly reserved matters may in the future be devolved.

This part of the Act was the subject of numerous amendments. In addition to the above provisions, clause 28 of the Bill prescribed in detail how the question of over-lapping powers was to be dealt with. It gave to the Scottish Parliament the power to legislate for Scots private law or Scots criminal law in an area reserved to the UK Parliament provided that the law would apply consistently to devolved and reserved matters. The legislative competence of the Scottish Parliament was protected against 'accidental' encroachment by giving it the power to amend UK legislation on a reserved matter where such legislation was consequential or incidental to any Scottish legislation in a devolved area. Furthermore, clause 28(5) provided that a matter should not become a reserved matter 'incidentally' by virtue of a UK enactment. Clause 28(8) included a presumption that an Act of the Scottish Parliament was within its legislative competence and was to have effect

19 The exact legal status of Acts of the Scottish Parliament is not made clear in the Act As discussed below, they are deemed to be subordinate legislation for the purposes of the Human Rights Act 1998. Within the Scottish legal order however they will have the same effects as acts of the UK Parliament previously had An Act of the Scottish Parliament on, for example, housing will have the same effect as previous legislation on this subject. The language of the Scotland Act, the use of terms such as Bill, Act, Royal Assent suggests that Acts of the two parliaments are equivalent. It can be argued that UK constitutional law is not yet sufficiently sophisticated to encompass these issues.

 © The Modern Law Review Limited 1999

accordingly. This prescriptive detail has been simplified by providing that the question of whether a provision relates to a reserved matter is to be determined by reference to its purpose having regard to its effect in all the circumstances. However a provision relating to a devolved matter which modifies Scots private or criminal law in a reserved area is now to be considered a reserved matter unless its purpose is to make the law apply consistently. In this formulation the pendulum seems to have swung in favour of Westminster but it will be for the courts to determine the precise boundary when devolution issues are raised in particular cases.

It would appear from all these provisions that there is a line, albeit fluid, defining the respective legislative competences of the Scottish Parliament and the UK Parliament. Overshadowing these provisions, however, is section 28(7) which states that the power of the Parliament of the UK to make laws for Scotland is unaffected by the devolution of legislative powers to the Scottish Parliament. In principle, therefore, the UK Parliament could continue to legislate for Scotland even in devolved matters. Should the UK Parliament choose to exercise these powers then the courts may find themselves faced with two conflicting legislative provisions emanating from two different legislatures. Ultimately the Judicial Committee of the Privy Council will be left to determine the validity of a particular rule[20] and to resolve what might be a constitutional crisis.

Section 28(7) highlights one of the weaknesses of the Scotland Act and, indeed, of the entire devolution process within the UK. The process of modernising the constitution is taking place in the context of a deeply conservative view of the nature of the UK constitution, one that relies heavily on an increasingly contested theory of the sovereignty of the UK Parliament. Within this traditional and conservative theory section 28(7) is easily explicable. The UK Parliament is sovereign and the Scottish Parliament is subordinate. The White Paper had indicated that this was to be the approach taken in the legislation. The Scottish Parliament is not to be seen as a reflection of the settled will of the people of Scotland or of popular sovereignty but as a reflection of its subordination to a higher legal authority. Following the logic of this argument, the power of the Scottish Parliament to legislate can be withdrawn or overridden. No attempt is made within the Act to specify the circumstances in which the UK Parliament might exercise these powers to legislate nor to specify any procedural rules of 'manner or form' that might be followed. This reliance on traditional constitutional theory in New Labour's attempt to modernise the constitution is disappointing. On the one hand New Labour looks to the decentralisation of power as a means to bring the UK constitution into the new millennium. On the other hand, it remains rooted in a conservative, indeed imperialist, past. The Scotland Act does not therefore recognise that devolution has an autochthonous nature; it is instead a reflection of the hegemony of one set of institutions over another.

Pre-legislative scrutiny

Responsibility for ensuring that the Scottish Parliament acts within the limits of its legislative competence rests with members of the Scottish Executive, the Parliament's Presiding Officer, the Law Officers, the Judicial Committee of the Privy Council and the Secretary of State for Scotland. A Scottish Bill is under continuous scrutiny as to its vires 'on or before' introduction into the Parliament

20 See the section below on devolution issues, 257–259.

© The Modern Law Review Limited 1999

(section 31). At this stage, the member of the Scottish Executive responsible for the Bill must state that its provisions are within the legislative competence of the Parliament and the Presiding Officer must decide whether its provisions are within the Parliament's legislative competence and state his/her decision.[21] At any stage in the legislative process and up to four weeks after the completion of the parliamentary procedures, the Advocate General,[22] the Lord Advocate or the Attorney General may submit the Bill to the Judicial Committee of the Privy Council for a decision as to whether any provision would fall outside the legislative competence of the Parliament.[23] Finally, the Secretary of State for Scotland during the same period can intervene by making an Order, stating the reasons for the Order, addressed to the Presiding Officer prohibiting him/her from submitting a Bill for Royal Assent if the Secretary of State believes that the Bill contravenes an international obligation or if it may have an adverse effect on the operation of UK law in an area reserved to the UK Parliament (section 35).

No doubt this belts and braces approach is designed to ensure that the new Parliament does not attempt to flex its muscles deliberately or inadvertently and stray over into the territory reserved to the UK Parliament. It is an excessively cautious approach suggesting an absence of trust on the part of the UK Parliament. It might also be argued that the arrangements place too much power in the hands of the Executive who will be able to control the activities of the Parliament.

Legislative procedures

The Scottish Parliament is unicameral and therefore the legislative procedures will differ from those used at Westminster. However the Scotland Act tells us very little about the legislative process. Section 36 provides that standing orders will be adopted which must ensure that a Bill receives three readings prior to its adoption. The three stage proceedings are similar to those used in the Westminster Parliament. The first stage is a general debate on the principles of the Bill followed by a second stage of more detailed scrutiny of its provisions. The final stage is to determine whether the Bill is accepted or rejected. Standing orders may also be adopted to provide for an expedited procedure for a particular Bill or for different procedures for Bills that restate the law, repeal spent enactments or for Private Bills. Only in limited circumstances may the Parliament reconsider a Bill after its passing. Once a Bill is accepted, the Presiding Officer will forward it to the Queen for Royal Assent. Precisely how the legislative procedures will operate is for the Parliament to determine.[24]

---

21 The Bill had originally sought to protect the position of the Presiding Officer by allowing the Parliament to overturn his/her decision on this matter.

22 The Advocate General for Scotland is a new post created by section 87. Given that the Lord Advocate and the Solicitor General for Scotland become part of the Scottish Executive after devolution, a new post is required in the UK government to advise on matters relating to Scots law and Scottish issues

23 A very curious provision was inserted in the late stages of the passage of the Act. Section 34 provides that if the Parliament resolves to reconsider a Bill that one of the Law Officers has referred to the Judicial Committee which has in turn referred a question for a preliminary reference to the European Court of Justice then the initial reference to the Judicial Committee may be withdrawn. This provision is nonsensical in terms of European law since the European Court of Justice does not have jurisdiction to rule on such hypothetical issues as whether proposed legislation might be contrary to European law if it were to be enacted.

24 The Act is silent on how Scottish Bills in areas reserved to Westminster are to be dealt with. The role of and the need for Scottish committees in Westminster are currently under scrutiny in Westminster

© The Modern Law Review Limited 1999

Standing orders and working methods

One of the most significant yet most succinct provisions in the Scotland Act is section 22 which states that 'the Parliament shall be regulated by standing orders'. The Act specifically requires standing orders to be adopted for the registration of members' interests and creates an offence of failure to comply with any provisions for registration and declaration of such interests (section 39). Schedule 3 states that standing orders must provide ways for maintaining order, for ensuring that its proceedings are public, for publication of its proceedings, certain aspects of the work of committees and certain Crown interests. As stated above, standing orders will also determine the exact nature of the legislative proceedings. It is for the Parliament to determine its standing orders after the elections have taken place. However, in preparation for the new Parliament the government decided to set up a Consultative Steering Group (hereafter CSG) under the chairmanship of Henry McLeish, Minister responsible for devolution, to report to the Secretary of State by the end of 1998. This report will be available for the new Parliament to accept or reject as it sees fit.[25]

The importance of standing orders for the new Parliament cannot be overestimated, as the way in which the Parliament will work will determine its character. The vision of the Scottish Constitutional Convention was of a Parliament characterised by 'accountability, accessibility, openness and responsiveness'. These principles have been adopted by the CSG. In the discussion papers presented to the CSG, at conferences organised in Scotland to discuss the workings of the new Parliament and in public debate there is a palpable desire that the new Parliament should be very different from Westminster. Crick and Miller have argued for a clean break with Westminster in many significant respects. They argue for a less confrontational and more consensual approach which would be 'better suited to and arising from Scotland's more democratic civic traditions'.[26]

At the time of writing, the final report of the CSG was not available. However some information is available about the kind of proposals that will be made.[27] The new Parliament is likely to be more informal than Westminster. MSPs will address each other by name. There will be no elaborate and formal rituals. Normal working hours will be observed. There is likely to be a weekly printed report and an overnight Internet report of proceedings. The most significant recommendations will be in the use of committees within the new Parliament. The Parliament will probably conduct the bulk of its business in committee. Unlike Westminster there is unlikely to be a

25 The CSG has attempted to use open and transparent procedures throughout the course of its deliberations. The CSG is worthy of a study in its own right. Its minutes, agendas and working papers can be found on the Scottish Office devolution web site, n 11 above. Of considerable interest too is the work of the Scottish Council Foundation, an organisation set up in 1997 under the auspices of the Scottish Council for Development and Industry to 'contribute towards the better government of Scotland'. For information see http://www.scottishpolicynet.org.uk/scf/about/about.htm.

26 Bernard Crick and David Millar, *To Make the Parliament of Scotland, A Model for Democracy* (Edinburgh, John Wheatley Centre, 1997) 3. Note the John Wheatley Centre is now named the Centre for Scottish Public Policy.

27 Information in this and the following two paragraphs is based on speeches given by George Reid and David Millar to a conference on The Scottish Parliament and the People held in Perth 22 October 1998. Summaries of the papers given at the conference have been published by the Centre for Scottish Public Policy. For further information see the following papers of the CSG: CSG (98) 9 on the working pattern of the Parliament, CSG (98) 19 on the operation, role and structure of committees of the Scottish Parliament, CSG (98) 20 on the Scottish Parliament and pre-legislative scrutiny, CSG (98) 21 on the Scottish Parliament: time-tabling of business, CSG (98) 22 on options for reporting and publishing proceedings, CSG (98) 40 on the role of rapporteurs, CSG (98) 45 on CSG decisions taken to date (July 1998), CSG (98) 45 on the pattern of business of the Scottish Parliament, CSG (98) 63 Bills-vires issues. All CSG papers are available on the Scottish Office devolution web site. The final report, published in January 1999, confirms these points. See *Shaping Scotland's Parliament* (HMSO, 1998).

© The Modern Law Review Limited 1999

distinction between standing and select committees with powerful all-purpose committees allowing MSPs to develop expertise in certain areas. The committees will be involved at all stages of the legislative process from the pre-legislative phase when the need for legislation might be discussed through legislative scrutiny of detailed proposals and with follow up debates on the implementation and enforcement of the legislation. In addition, the possibility that committees might initiate legislation is being discussed. Committees would have a strong role to play in ensuring accountability of the Scottish Executive. The Act lays down the circumstances under which Parliament and its committees may exercise powers to require Scottish Ministers, civil servants and others to attend to give evidence or to provide documentation across the whole range of devolved matters. In limited circumstances, Ministers of the Crown might also be required to attend. Evidence may be required under oath. Failure to give evidence will be an offence (sections 23–26).

The CSG is likely to recommend the creation of a Bureau of Business Committee. This type of committee, which is well known in Europe, is not used in Westminster. The Business Committee will be composed of the Presiding Officer, representatives of the political parties, and possibly the convenors of the other main committees. The Bureau would be responsible for settling the work of the committees and the Parliament, to prepare an annual programme of work, to deal with timetabling of debates and other business of the Parliament. The Bureau would be a very powerful committee and therefore its decisions would need to be scrutinised by the whole House. Perhaps a vote on its report would be needed.

The Presiding Officer

If a Business Committee is to be created to organise the business of Parliament then the person who chairs this committee will have a crucial role in defining how the new Parliament will conduct its business. The Presiding Officer will take the chair. A Presiding Officer is to be elected at the first meeting of Parliament and he/she will remain in office until the next election for a new Presiding Officer or until he/she resigns or loses his/her seat in Parliament (section 19). The Act also provides for the election of two Deputy Presiding Officers. Elsewhere in the Act, the Presiding Officer is given specific tasks in determining questions of vires and in the legislative process, for example in conveying a Bill for Royal Assent. However the Act gives little flavour of the extent and complexity of the Presiding Officer's role. In its discussions the CSG have examined the role of the Presiding Officer in a wider context. He/she will, for example, represent the Parliament in interactions with the Scottish Administration, the UK Parliament and the other devolved assemblies and represent the Scottish Parliament in interactions with other Parliaments outside the UK.[28] The CSG does not envisage a role that is exactly comparable to the role of the Speaker of the House of Commons but there are parallels with this role.

The Scottish Administration

The Scottish Executive

Sections 44 to 63 of the Scotland Act set out the structure and functions of the Scottish Administration. In several respects the relationship between the Scottish

28 CSG (98) 78.

© The Modern Law Review Limited 1999

Executive and the Scottish Parliament mirrors that between the UK government and the UK Parliament. In Scotland these relationships are formalised by statute rather than convention and in this sense the Scotland Act looks very much like a written constitution for Scotland. The Act creates a Scottish Executive composed of a First Minister, Ministers appointed by the First Minister and the Lord Advocate and Solicitor General for Scotland. Collectively these are known as the Scottish Ministers (section 44). The First Minister is to be appointed by Her Majesty from among the MSPs and he/she must command the confidence of Parliament. He/she must resign on a vote of no confidence. If the office of First Minister is vacant or if the First Minister is unable to act, the Presiding Officer of the Parliament may designate another MSP to act as First Minister. The First Minister is the keeper of the Scottish Seal (Section 45). It is the function of the Parliament to nominate the First Minister and it must do so within 28 days of an election or the post of First Minister becoming vacant. The Presiding Officer of the Parliament nominates the First Minister to Her Majesty (Section 46).

Once appointed, the First Minister will appoint Ministers from amongst the ranks of MSPs. A person who holds a ministerial office at Westminster may not be appointed as a member of the Scottish Executive and vice versa. The First Minister must obtain the approval of Parliament for such appointments.[29] Ministers hold office at Her Majesty's pleasure and may be removed by the First Minister. They may resign their ministerial posts at any time and must do so when Parliament approves a vote of no confidence. On resignation a Minister ceases to hold office immediately. The First Minister may also appoint junior Scottish Ministers to assist Scottish Ministers in the exercise of their functions, Such junior Scottish Ministers are subject to the same rules as apply to Scottish Ministers (section 49). The Act does not specify the number of Scottish ministers but leaves this to the decision of the First Minister and the Parliament. Henry McLeish, Minister for Devolution, is reported to favour a small Cabinet in line with the current Scottish Office team whereas Alex Salmond has nominated 20 names for his potential Cabinet.[30] No provision is made for Ministers or junior Ministers to be appointed from outwith the ranks of MSPs. The recent example of the appointment of Gus MacDonald to the Scottish Office and his simultaneous elevation to a peerage cannot therefore be repeated in the Scottish Executive.

The First Minister also recommends the appointment or removal of persons to the posts of Lord Advocate and Solicitor General for Scotland who, after the Scotland Act comes into force, will cease to be Ministers of the Crown (section 48). The Lord Advocate retains his/her independence as head of the system of criminal prosecution and investigation of deaths in Scotland. The Lord Advocate and the Solicitor General need not be MSPs but they have the right to participate in the proceedings of the Parliament although they may not vote (section 27). The first Minister will also recommend to the Prime Minister the persons to be appointed as Lord President of the Court of Session and Lord Justice-Clerk. He/she will recommend to Her Majesty the appointment of other judges in Scotland. More controversially the Bill would have allowed the First Minister to recommend removal of judges following a resolution of the Scottish Parliament. After criticism

29 This position differs from that at Westminster. See R. Brazier, 'The Scottish government' (1998) *Public Law* 212 for the view that 'this is a strong assertion of parliamentary rights in relation to the executive which Westminster should (but no doubt will not) copy.' It is comparable to the role of the European Parliament in the appointment of members of the European Commission. It will be interesting to see if the Scottish Parliament will conduct hearings in the same way as the European Parliament has done before giving its approval to appointments.

30 *The Herald* 22 October 1998.

© The Modern Law Review Limited 1999

from the Scottish judiciary,[31] this was amended to provide for the creation of an independent tribunal to investigate judges who were deemed unfit. On the advice of the tribunal the First Minister can recommend the removal of a judge after the adoption of a resolution of the Parliament (section 95).

No provision was made in the Bill for an opposition. This was probably a reflection of the desire to have a more consensual approach to politics. However until elections take place it is not clear what will be the respective strengths of the political parties in the new Parliament and hence which party or parties are likely to be able to have sufficient support to command the confidence of the House. If two parties dominate the political scene then it is at least imaginable that the Parliament will divide into two camps of government and opposition. Section 97 now sets out the possibility for financial assistance to be given to opposition political parties.

Functions

Statutory functions may be conferred on the Scottish Ministers. These functions are exercisable on behalf of Her Majesty and may be exercised by any member of the Scottish Executive.[32] The Scottish Ministers share joint responsibility and liability for their acts or omissions except where functions are conferred on the First Minister alone or where the Lord Advocate retains certain functions (section 52).

In areas of devolved competence functions are transferred to the Scottish Ministers from Ministers of the Crown. These include functions relating to the exercise of the prerogative and to functions conferred by any pre-commencement Act or enactment (section 53). The Scottish Minister may not make any provisions by subordinate legislation that would be outwith the legislative competence of the Parliament. Neither may they carry out any functions under any Act of the Scottish Parliament that are outwith the legislative competence of the former. Where competence has been transferred and where a provision requires consultation with another Minister of the Crown then that requirement is not to apply except in the one specified area of the designation of enterprise zones (section 55).

Certain powers are to be shared by Ministers of the Crown and by the Scottish Ministers. These areas are listed in section 56. A further shared power is in relation to implementation and enforcement of obligations under European Community law. In this respect Ministers of the Crown retain the competence to adopt subordinate legislation for Scotland even in devolved areas. These powers are in addition to the powers of Scottish Ministers to introduce such delegated legislation (section 57). Furthermore, the Act provides that no member of the Scottish Executive may make subordinate legislation that is incompatible with European Community Law or with a Convention right.

The Secretary of State may order that a proposed action (including the making or confirming of subordinate legislation) by any member of the Scottish Executive is incompatible with any international agreement and therefore that action should not be taken. He/she may also order that any action (including introducing a Bill to the Scottish Parliament) must be taken in order to ensure compliance with an international obligation. The Secretary of State may also, after stating his/her

31 Lord McCluskey described the original clause 89 as being 'an instrument of the kind that we find in every tin pot dictatorship and every banana republic where those in power do not want to be troubled by an independent judiciary'. *The Herald* 6 October 1998.

32 The Scottish Parliament must make provision for investigation of complaints against members of the Scottish Administration or Executive in carrying out their functions (section 91).

© The Modern Law Review Limited 1999

reasons, revoke subordinate legislation which would be incompatible with any international obligation or UK defence or security interest or if it is within a devolved competence but its operation would have an adverse impact on the operation of any UK enactment (section 58). Additional powers may be transferred to Scottish Ministers by Order in Council. These powers may be exercised by Scottish Ministers instead of a Minister of the Crown, concurrently with such a Minister or by a Minister of the Crown after consultation with or agreement of the Scottish Ministers (section 63). Conversely section 108 provides for a transfer of functions from Scottish Ministers in the same way.

The distinction between devolved powers, shared powers and powers exercised after consultation with a Minister of the Crown is not completely clear within the terms of the legislation. There are likely to be disputes over which ministers have competence over a range of matters. A good example of this confusion lies in the handling of European business. The UK is responsible at European level, including before the European Court, for any failure to respect European law by Scottish Ministers or Parliament.[33] Matters relating to the European Community are reserved but responsibility for implementation and enforcement of Community obligations is not reserved. Responsibility for implementing Community obligations in devolved areas lies with the Scottish Ministers and Parliament but this function is shared with Ministers of the Crown. The Scotland Act therefore protects the position of the UK in three ways. The first is by prohibiting the Scottish Executive from infringing Community Law. The second is by providing a concurrent power to introduce subordinate legislation to comply with Community law and the third is the overarching authority given to the Secretary of State to require or prohibit particular legislation. Using one or other of these mechanisms, it is possible to ensure that Community law is observed within Scotland.

Some mechanism is undoubtedly needed to ensure that Community law is observed in Scotland. The granting of concurrent powers to Ministers of the Crown does, however smack of a rather paternalistic approach. It is not clear in the legislation when Ministers of the Crown could exercise their powers to issue subordinate legislation covering Scotland nor what the courts should do if they are faced with different approaches to implementation of a Community obligation coming from a Scottish Minister or from a Minister of the Crown. In order to iron out such difficulties, Ministers and officials are in the process of negotiating a Concordat defining the role of the Scottish Executive in European matters. The Concordat is intended to deal with such questions as negotiating the UK position on EU matters, deciding who will attend meetings of the Council of Ministers, resolving differences of view, implementing EU obligations and handling infraction proceedings.[34]

33 Case C-33/90 *Re Toxic Waste in Campania: Commission* v *Italy* [1992] ECR 1573. The European Court stated that 'although each member state is free to distribute legislative powers as it wishes in the internal respect, it nevertheless remains responsible to the Community for fulfilling obligations under Community law'.

34 CSG (98) 74. Concordats are being developed on financial assistance and on the exchange of statistical information as well as on EU matters. Concordats are not intended to have legal effects and are not 'wholly prescriptive' but are 'the basis for dialogue'. They may be seen as the beginnings of written constitutional conventions or as political cohabitation arrangements. They will operate alongside a joint ministerial committee, an advisory body on matters of communal interest. See HL Deb vol 592 col 1488 28 July 1998 and the Select Committee on Scottish Affairs, 2nd Report 1998, *The Operation of Multi-Layer Democracy* para 41.

© The Modern Law Review Limited 1999

The role of the Secretary of State for Scotland

The Secretary of State for Scotland is a Minister of the Crown and will continue to be so after the Scotland Act comes into force. The Act does not define the future role of the Secretary of State but it is clear that such a post will need to continue albeit with diminished functions. Crick and Millar argue that a Secretary of State will be needed to exercise 'the reserved powers as they affect Scotland' and to be 'the United Kingdom's negotiator and adviser on relations with the Scottish Government and Parliament'.[35] The White Paper echoes this view stating that the first task of the Secretary of State is to ensure that the legislation establishing the Parliament is secured and to support the initial development of the new institutions. Thereafter the focus will be on promoting communication between the Scottish Parliament and Executive and between the UK Parliament and Government on matters of mutual interest; and on representing Scottish interests in reserved areas'. The Secretary of State is, however, given certain defined functions by the Act in relation to establishing the vires of an Act of the Scottish Parliament and in relation to ensuring that the Scottish Executive does not infringe any international obligations of the UK. The distinction between the Secretary of State and the Scottish Ministers is also crucial in matters relating to Scottish finance. There are suggestions that in the future the role of Secretary of State for Scotland might be taken over by a minister for Constitutional Affairs with more general duties.

Financial provisions and tax varying powers

The bulk of Scottish financing after devolution will continue to come from the United Kingdom budget under the 'block and formula' system. Until now responsibility for spending has been in the hands of the Scottish office but after devolution budgetary responsibility will lie with the Scottish Parliament. The Scotland Act creates a Scottish Consolidated Fund, to be held by the Paymaster General, into which the Secretary of State will make payments out of money provided by the UK Parliament. Any other sums received by the Scottish Administration will also be paid into this Fund. When required, the Scottish Minister may make payments to the Secretary of State out of the Fund (section 64). Sums can only be paid out of the fund when authorised by an enactment of the Scottish Parliament or under the terms of the Scotland Act itself, for example to meet the costs of the Parliament and the Scottish Administration (section 65). Scottish Ministers are empowered to borrow money from the Secretary of State for certain limited purposes. Such funds are to be provided from the Treasury. Scottish Ministers are also empowered to lend to Scottish public bodies on terms specified within the Act (sections 66–68). These financial provisions should be read in the light of the information provided by the Treasury in December 1997. The Treasury set out the principles under which the block grant would operate after devolution.[36]

The most crucial aspect of all the financial arrangements lies in section 70 that requires the Scottish Parliament to make rules in respect of the budget. It must make arrangements to ensure that Scottish Ministers, the Lord Advocate and others prepare proper accounts in respect of sums paid to them from the Scottish Consolidated Fund. It must require that Scottish Ministers prepare accounts of

35 n 26 above.

36 HM Treasury news release 157/97 of 8 December 1997 *Principles to govern Scotland's and Wales's shares of public spending under devolution.*

© The Modern Law Review Limited 1999

sums paid into the Fund. Access to documents must be guaranteed for those responsible for preparing such accounts. Members of the Scottish Administration must be designated whose task is to be answerable to Parliament in respect of the expenditure and receipts of the Scottish Administration. Arrangements must be made for the publication of accounts and their presentation to Parliament. The Act creates the post of Auditor General for Scotland who must be nominated by the Parliament but who must be independent of the Parliament and the Scottish Executive. He/she will be appointed by Her Majesty. The Auditor General can only be removed from office after a resolution of the Parliament with a two-thirds majority in favour. His/her functions are listed in the Act and include issuing credits for payments, examining parliamentary accounts and examining the economy, efficiency and effectiveness of public spending.

A Financial Issues Advisory Group (FIAG) was established to report to the CSG on proposals for legislation and procedures on financial and budgetary matters. At the time of writing the final report of that group is not available although a preliminary report was made available in July 1998.[37] FIAG is likely to recommend the creation of a budgetary committee as well as an audit committee and that the new Parliament must approve spending based on estimates prepared by the Scottish Executive. FIAG supports openness in budgetary matters so as to enhance accountability for public spending in Scotland. To this end it recommends that any proposed measure should be costed and scrutinised prior to Parliament approving measures.

After the tax year 2000–2001 the Scottish Parliament, on a motion from a member of the Scottish Executive, has the power to adopt a tax varying resolution. Such a resolution can increase or decrease the basic UK rate of income tax for Scottish taxpayers by three pence. It can only apply for one tax year and a further resolution would be needed for any subsequent tax year. No tax is to be levied on savings or dividend income. The Act details the effects which the timing of such a resolution would have as well as defining who is a Scottish taxpayer and, therefore, subject to such a tax varying resolution (sections 73–75). There is an attempt to protect these provisions against any changes which might be made to the Income Tax Acts and which have the effect of altering the impact of the tax varying power. Any proposals to amend the Income Tax Acts which would affect the tax varying power must be laid before the House of Commons by the Treasury explaining why the tax varying powers need to be amended. The effect of any such amendments should, in so far as possible, secure the practical extent of the Scottish Parliament's tax varying powers.

The Act makes provision for any additional income tax paid by Scottish taxpayers to be paid into the Scottish Consolidated Fund. Where the tax varying power of the Scottish Parliament is used to decrease the rate of income tax, any overpayments must be repaid by the Scottish Ministers from the Fund.

Devolution issues

Despite the numerous checks on the Scottish Parliament and the Scottish Executive in the Scotland Act which serve to ensure that each acts within the limits of its devolved powers, there will inevitably be disputes about where exactly the border lies between devolved and reserved powers. The procedures for resolving these constitutional issues, or what the Act terms devolution issues, are contained in Schedule 6 and are positively Byzantine.

37 CSG (98) 10 and CSG (98) 37.

© The Modern Law Review Limited 1999

The issues that might arise are listed in the Schedule. Whether a matter is within the legislative competence of the Scottish Parliament, whether subordinate legislation is within the ambit of devolved powers, whether the exercise of functions by the Scottish Executive is a reserved matter, whether subordinate legislation or the exercise of particular functions is, or would be, incompatible with Community law or with the European Convention on Human Rights, whether the failure to do something is a similar violation or whether a matter in which a Minister of the Crown has purported to act is a reserved or devolved matter are all devolution issues. Such issues fall to be resolved by the courts. Provision is therefore made for each and every court or tribunal in the UK as to how to handle such issues. Schedule 6 provides for a complex scheme of references and/or appeal mechanisms on devolution issues with ultimate authority being vested in the Judicial Committee of the Privy Council.[38] Essentially, lower courts may refer devolution issues (and must if it is a court or tribunal from which there is no appeal) to the higher courts. These courts may refer matters to the Judicial Committee of the Privy Council which is also, in some circumstances, the court of appeal from the higher courts. The Law Officers have a right both to intervene in cases and to institute proceedings and they may require a case to be referred to the Judicial Committee.

Questions of interpretation are governed by section 101. Provisions of Acts of the Scottish Parliament or subordinate legislation that might be read as being outside competence are to be read 'as narrowly as is required for [it] to be within competence'. If such a reading is possible then the provision is to have effect. This approach contrasts sharply with that in the Bill where there was an instruction to read any provision, as far as possible, 'to be within the legislative competence of the Parliament'.[39]

The part of the Act dealing with devolution issues should be read in conjunction with the Human Rights Act. For the purposes of the Human Rights Act, Acts of the Scottish Parliament are not treated in the same way as Acts of the UK Parliament even when the latter concern purely English matters. Acts of the Scottish Parliament are defined as subordinate legislation. The Human Rights Act requires the courts to read this legislation in such a way as to render it compatible with Convention rights. Where it is incompatible with the terms of the Human Rights Act or a Convention right then the courts must provide the appropriate remedy. In an application for judicial review, the Human Rights Act states that in Scotland an individual must have title and interest to sue and bring proceedings within the defined time-scale. The Scotland Act states that an individual may only bring proceedings against an act, defined as legislation or action or failure to act by the Scottish Executive, if he/she would be a victim of that violation as defined in the European Convention and the Human Rights Act. Where a human rights issue is also a devolution issue the mechanisms for appeal and reference in Schedule 6 will apply but where a human rights issue is not a devolution issue then the provisions of the Human Rights Act alone will apply. Perhaps the most significant practical aspects of this is in relation to the need to exhaust domestic remedies before an

38 Space precludes a detailed description of Schedule 6. See J. McFadden, 'The Scottish Parliament: Provisions for Dispute Resolution' (1998) *Juridical Review* 221, T. Jones, 'Scottish Devolution and Demarcation Disputes' (1997) *Public Law* 283. On the reasoning behind the choice of the Judicial Committee see the essay by the current Solicitor General for Scotland (written before he assumed that role), C. Boyd, 'Parliament and Courts: Powers and Dispute Resolution' in T. StJ. Bates (ed) *Devolution to Scotland: The Legal Aspects* (Edinburgh. T. & T Clark, 1997).

39 Clause 28 (9).

© The Modern Law Review Limited 1999

individual can take his/her case to the Convention authorities. In criminal cases involving a devolution issue an additional level of appeal has been introduced in the form of the Judicial Committee.

Similar considerations arise in relation to European Community law matters in Scotland. Civil cases raising questions of interpretation and application of Community law have traditionally gone to the House of Lords for resolution or possible reference to the European Court of Justice.[40] There is now the possibility that these matters, where they relate to Acts of the Scottish Parliament or acts of the Scottish Executive, will be defined as devolution issues and, as such, may be referred to the Judicial Committee for decision or for possible reference to the European Court of Justice. This might occur even after they have gone to the House of Lords if the devolution issue is raised only at that late stage. From the point of view of a uniform interpretation and application of Community law throughout the UK, and indeed within Scotland, this may not be the best solution although before cases begin to arise it is difficult to predict how the situation will unfold. There is certainly the possibility that on matters relating to Community obligations implemented at UK level within Scotland that the House of Lords remains the supreme court but on matters of Community law implemented and enforced at the Scottish level that the Judicial Committee becomes the supreme court. The possible confusion is undesirable.

The Scotland Act contains few provisions relating to the organisation or functions of the Judicial Committee when it meets to decide devolution issues. Section 102 requires the Judicial Committee to meet in open court and provides that the decisions of the Judicial Committee shall bind all other courts – including the House of Lords. Its members, when it sits as a 'devolution court', must hold or have held office of a Lord of Appeal in Ordinary or a high judicial office as defined in section 25 of the Appellate Jurisdiction Act 1878. There is no requirement that members of the Judicial Committee be drawn from any particular jurisdiction within the UK although Orders in Council are to be adopted to confer any relevant powers on the Judicial Committee and to regulate its proceedings.

In many respects, this section of the Act is the least satisfactory. Little thought or planning appears to have been given to any alternative to the Judicial Committee yet it seems to have been given the de facto role of a UK constitutional court without considering in full the implications of that decision. For example the confusion discussed above in relation to human rights and European Community matters would not arise if these matters were all entrusted to a constitutional court. Also the use of the Judicial Committee and the appeal procedures laid down are inordinately complex. A simpler method would have been to provide a preliminary reference procedure, comparable to that used in references on questions of interpretation of European law to the European court, where specific devolution issues could be resolved. Again a constitutional court could serve this function. As it stands legal proceedings that by virtue of the appeal system will invariably be both lengthy and costly have the potential to hinder the work of the Scottish Parliament and the Scottish Executive.

40 The House of Lords traditionally has not had jurisdiction in criminal appeals coming from Scotland. This tradition comes from the nature of the old Scots Parliament which itself did not have jurisdiction in criminal appeals. Jurisdiction could not therefore be transferred at the time of the Acts of Union. See T.B Smith, *A Short Commentary on the Law of Scotland* (Edinburgh: W Green & Son Ltd, 1962) 87. The Act requires devolution issues arising in the House of Lords to be referred to the Judicial Committee unless the House of Lords considers it appropriate to determine the issue itself.

© The Modern Law Review Limited 1999

Conclusions

The Scotland Act was introduced by the Labour government to fulfil an election promise that it would continue to support in Scotland a policy for home rule that it had encouraged and developed whilst in opposition. It was introduced in the context of a wider programme of modernisation of the constitution. The success of the devolution process in Scotland will eventually therefore be measured against both these objectives. Does the Scotland Act answer the aspirations of the people of Scotland and does it contribute to the process of modernisation? The answer to both these questions at present must be a qualified yes. The Scotland Act creates the structures of Scottish home rule. In particular it creates a Parliament with extensive legislative powers. It also provides the possibility of an extension of these powers so that devolution is not a 'one off' transfer of powers but a process that will require a continuing interaction between Holyrood and Westminster, Edinburgh and Whitehall. To some extent the Scotland Act dictates the way in which the new Scottish institutions will operate but it leaves most aspects of the working of the institutions to develop over time. Non-statutory mechanisms will need to be developed governing more precisely the relationships between the Scottish Parliament and local authorities in Scotland,[41] between Scottish institutions and UK institutions, and between the devolved regional institutions themselves. The Act itself is silent on these matters. These relationships will be developed by political agreements (or disagreements) leading to the formation of new constitutional conventions and law. It is possible to discern already the development of a new constitutional language of concordats and covenants. The courts too will need to develop new constitutional skills in determining the precise boundaries between devolved and reserved matters and in developing the tests to be used in resolving devolution issues.

Whether the Scotland Act answers the aspirations of the Scottish people will depend very much on how demands for home rule change. The Scotland Act, by and large, answers the demand of the Scottish Constitutional Convention. These demands were fashioned without any real input from the Scottish Nationalist Party and the aspirations of the SNP and its supporters are not for devolution but for independence. It is impossible to tell how widespread such support is.

The Scotland Act does contribute to the process of modernisation. Its particular importance lies in the recognition that the United Kingdom is a union, not a unitary, state. The creation of separate institutions of government brings Scottish political institutions into line with Scottish civil society, operating in the context of the union. The Act itself, however, does not modernise the UK's constitutional order. The reliance on traditional theories of the sovereignty of the UK Parliament and of the relationship between sovereign and subordinate parliaments, the failure to recognise the need for a constitutional court, the failure to recognise the legal personality of Scotland within the UK's constitutional structure all point to a conservative rather than a modern approach to constitutional reform. That said, the Scotland Act, taken together with other aspects of the government's legislative programme, can be seen as the first step towards a modernisation of the UK constitution. At the very least it presents constitutional lawyers with an opportunity to re-evaluate many aspects of constitutional law and theory to try to explain the new complexities of government in the UK.

41 A committee has been established under the Chair of Professor MacIntosh to examine the relations between the Scottish Parliament and Scottish local authorities after devolution. At the time of writing its report has not been completed although early press reports point to the need for a 'covenant' between the Scottish Parliament and local authorities to protect the latter against encroachment from the former.

© The Modern Law Review Limited 1999

Detecting, Prosecuting and Punishing Benefit Fraud: The Social Security Administration (Fraud) Act 1997

*Gráinne McKeever**

Social security fraud is currently estimated at £7 billion per year out of an annual social security budget of £90 billion.[1] The social security system is hugely bureaucratic and administrative, leaving it wide open to abuse. Obviously there are individuals keen to exploit such vulnerability, but many innocent claimants also fall foul of its rules, quite inadvertently. In 1997, the Conservative government declared that its anti-fraud initiatives would save taxpayers £7 billion over the next three years.[2] The main plank in the Conservatives' anti-fraud strategy, which has now been implemented by the Labour Government, was the Social Security Administration (Fraud) Act 1997, designed to make the detection, prosecution and punishment of benefit fraud more effective.[3] The aim of this note is to examine the legislation previously used in the battle against benefit fraud, to identify the rationale for legislative change, and to analyse the problems which arise from this new legislative position.

Social security fraud prior to the 1997 Act

Social security fraud was established as a criminal offence prior to the enactment of the Social Security Administration (Fraud) (SSAF) Act 1997. Section 112 of the Social Security Administration Act 1992 provides that claimants can be prosecuted if they are shown to have made a statement or representation which they know to be false, or if they have produced or have knowingly caused or allowed to be produced any documentation or information which they know to be false for the purpose of obtaining benefit for themselves or another, or for any other purpose connected with the legislation.[4] In the absence of a false statement or documentation, prosecution under the 1992 Act is not possible, but where it is both possible and successful it can result in a fine, imprisonment or both. Section 34 of the Jobseekers Act 1995 also makes benefit fraud an offence, although it is obviously limited to fraudulently obtaining Jobseeker's Allowance, with the scope of section 34 the same as that of section 112 of the 1992 Act. Most prosecutions for social security fraud are brought under social security legislation since this is comparatively cheaper than prosecuting under the criminal law.[5]

* School of Public Policy, Economics and Law, University of Ulster.

Thanks to Brice Dickson, University of Ulster, and Laura Lundy, Queen's University, Belfast, for their comments on earlier drafts of this note.

1 *Beating Fraud is Everyone's Business: Security the Future* Cm 4012 (1998) 12.

2 *ibid.*

3 The majority of the Act's provisions have been brought into force by the following commencement orders: No. 1 SI 1997 No 1577, No 2 SI 1997 No 2056, No 3 SI 1997 No 2417, No 4 SI 1997 No 2669 and No 5 SI 1997 No 2776.

4 The 1992 Act is a consolidating measure: s 112 of the Act was originally s 146(3)(c) of the Social Security Act 1975, which then became s 55 of the Social Security Act 1986.

5 Prosecutions under the Social Security Administration Act 1992 are 'relatively inexpensive ... about £150 a throw', while those brought under the criminal law cost over £1,000 per case: Social Security Committee, Third Report: *Housing Benefit Fraud*, 1995–96, HC 90-I and HC 90-II: volume 2, Q 704.

© The Modern Law Review Limited 1999 (MLR 62 2, March) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

However, social security fraud can also be prosecuted as theft or forgery. Under section 15 of the Theft Act 1968 a claimant who practises a deception as to either fact or law which results in him/her dishonestly obtaining benefit for him/herself or for another can be prosecuted for obtaining property by deception. The deception can be through words or conduct and is punishable whether it is deliberate or reckless. Cases brought under the Theft Act must be initiated by the police. However, without a clear admission of intent by the claimant, such a prosecution is unlikely to succeed and the police may therefore be unwilling to prosecute unless large sums of money are involved. Finally, prosecutions may also be brought under the Forgery and Counterfeiting Act 1981. A claimant can be prosecuted for forgery if, for example, he/she forges the signature on a benefit cheque, or forges an instrument of payment.

The rationale for change and the SSAF Act 1997

Detection

The detection of benefit fraud has, in the past, proved to be a problem. Data-matching has now been identified as the solution.[6] However, the Social Security Committee on Housing Benefit Fraud was given mixed evidence on whether effective data matching activities were possible under existing legislation or whether they were hindered by the Data Protection Act 1984.[7] Problems with the data protection legislation were raised by several witnesses who felt that the provisions were at best ambiguous and at worst protected the criminals rather than innocent claimants. The Inland Revenue identified the circumstances under which they felt permitted to disclose confidential information to the Department of Social Security (DSS), and these did not include disclosure for the purpose of fraud investigation.[8] Other witnesses before the Committee, however, refuted the suggestion that the Data Protection Act 1984 constituted a constraint, pointing to section 28(3)[9] which permits disclosure where fraud is suspected. In the light of this conflicting evidence the Committee recommended that if existing legislation could not be used, the Government should introduce legislation to enable confidentiality to be broken in cases of serious social security fraud.[10]

In its reply to the Committee's Third Report, the Government expressly stated that the data protection legislation already provided an exception for the transfer of information relating to fraud. Nevertheless, the SSAF Act introduces substantial changes. Under the Act data supplied to the tax authorities (section 1), government

6 Data matching is defined as 'the computerised comparison of two or more sets of records with the main objective of searching for records relating to the same individual': Social Security Committee, Third Report, n 5 above, volume 2, Appendix 4.

7 *ibid*, volume 1, paras 99–100. The 1984 Act is to be repealed by the Data Protection Act 1998, c 29. Minor provisions have been in force since 16 July 1998, with the remainder to be brought in at a date to be appointed.

8 *ibid*, volume 1 para 94.

9 Now s 29(3) of the 1998 Act. This provides that disclosure of personal data is permitted where it is for the prevention or detection of crime, the apprehension or prosecution of offenders, or the assessment or collection of any tax or duty, and where non-disclosure would be likely to hinder those objectives.

10 n 5 above, volume 1, para 96. Serious social security fraud is generally the result of organised fraud, the most common form of which is multiple applications for housing benefit, using false identities and forged documentation: *Local Government Chronicle*, 10 November 1995 17 Typical minor frauds include claimants failing to disclose their earnings to the Benefits Agency in order to remain eligible for Jobseeker's Allowance, or making a false declaration that they are living alone and financially unsupported.

© The Modern Law Review Limited 1999

departments (section 2) or authorities administering housing benefit or council tax benefit (section 3) may be supplied in turn to the Secretary of State for use in the prevention, detection, investigation or prosecution of offences relating to social security, or for checking the accuracy of information relating to social security. The central aim behind sections 1–3 is that any inconsistencies in information supplied to various government bodies will immediately be spotted so that suspected frauds can be investigated and uncovered. Obviously, this is to be welcomed, but a balance needs to be maintained between the public interest in defeating fraud and the duty of Parliament to safeguard civil liberties. The main criticism of the SSAF Act is that it fails to strike this balance.

The view of the Data Protection Registrar was that such extensive data matching powers require a corresponding increase in the legal protection of individual privacy, and that a statutory code of practice should be created for this purpose.[11] This view was echoed by the Select Committee on the Scrutiny of Delegated Powers,[12] but was ultimately rejected by the Government, which agreed to allow a code of practice to be drawn up but refused to give it statutory authority.[13] Under the Data Protection legislation, the Data Protection Registrar can serve an enforcement notice on a data user who has committed a criminal offence by breaching the principles of data protection enshrined in the legislation.[14] However, section 63(3) of the 1998 Act[15] states that a government department cannot be prosecuted for breach of those principles. With a non-statutory code and a Data Protection Act which has no teeth in this area, regulation of data matching practices by government departments becomes extremely weak, despite Government claims to the contrary.[16] The concern expressed by Lord Carter still remains, namely that '[i]f the Act stops the Government being prosecuted, how can the registrar enforce the provisions of [a] voluntary code?'[17] Data matching for the purposes of crime prevention can, if necessary, be defended under Article 8 of the European Convention on Human Rights (ECHR), which lays down the principle of respect for privacy, with such exceptions as are necessary in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others. However, the provisions in the SSAF Act extend beyond data matching for the prevention of crime and enable information to be passed between various bodies for the purposes of checking the accuracy of that information in relation to social security matters.

The Minister of State for the DSS, Lord Mackay, provided three reasons why the inclusion of such powers was important.[18] First, inconsistencies that are discovered through data matching may indicate that the data held on a particular person is inaccurate, and so there should be a corresponding power to correct that information.

11 HL Deb vol 579 col 169 11 March 1997.

12 *ibid.*

13 The government's view was that a non-statutory code could accommodate the anticipated implementation of the data protection Directive 95/46/EC. The code itself is to regulate areas such as the need to inform claimants that data matching may take place across government departments; the safeguards against wrongful accusations of fraud, the steps to be taken before benefit is suspended; and details of the complaints procedure: *ibid* col 177.

14 1998 Act, s 40 replacing 1984 Act, s 10(9).

15 Previously 1984 Act, s 38(2).

16 HL Deb vol 579 col 183 11 March 1997. Lord Mackay, Minister of State for Social Security, stated that any potential breaches of the data protection principles would, as in the past, be resolved between the Data Protection Registrar and the government department concerned to the satisfaction of the Registrar.

17 *ibid* col 185.

18 *ibid* cols 209–210.

© The Modern Law Review Limited 1999

However, the powers under the 1997 Act do not simply provide that data inconsistencies, once discovered, may be corrected. The provisions allow data matching to take place for the specific purpose of checking the accuracy of information. The two are very different, and while the former may be accepted as reasonable, justification for the latter needs to be specifically provided. Secondly, the Minister identified a legitimate public interest in ensuring that social security records are accurate, given the total public expenditure on social security. Nevertheless a balance is still required between the interests of society in maintaining accurate social security records and protecting the privacy of data held on a particular individual, a balance which Article 8 ECHR aims to achieve. The Data Protection Registrar has stressed the need for adequate safeguards to be put in place,[19] and in their absence, it is difficult to see how the necessary balance is achieved. Finally, it was argued that correction of inaccurate information enables social security authorities to comply with the fifth data protection principle set out in the 1984 Act[20] that personal data should be kept accurate and, where necessary, up to date. In many respects, this argument is the most valid of the three. Once again, however, the concern is that the power to cross check data for the purpose of assessing its accuracy is quite wide and, in the absence of a statutory code of practice, is one that is potentially very open to abuse. In 1990, Sir Nicholas Browne-Wilkinson warned of the need to protect the freedom of the individual, stating that '[t]he dossier of private information is the badge of the totalitarian state'.[21] Government measures taken to allay such fears may yet prove to be inadequate.[22]

Additional provisions in the 1997 Act to assist in the detection of benefit fraud include the power of the Secretary of State to initiate reviews of decisions relating to Disability Living Allowance, Disability Working Allowance and Attendance Allowance (section 17) and to require claimants to attend medical examinations in 'prescribed circumstances' for that purpose (section 18). Failure to attend the required medical examination may result in benefit being withheld. The main concern as regards these provisions is that a suspicion that the conditions of entitlement are no longer being met is not a prerequisite for initiating a review. Although the then Secretary of State did inform the House of Commons that the provisions were intended to review any 'suspect case'[23] nevertheless such a safeguard is not included in the wording of the legislation. A recent DSS Research Report on Social Security Fraud shows that fraud was *confirmed* in only five per cent of disability benefit cases, while *suspicion* of fraud (where the investigating officer strongly suspected fraud but had no proof) existed for even fewer cases.[24] The means adopted in the 1997 Act to deal with fraud potentially committed by those claiming disability benefits would therefore appear to be somewhat disproportionate to the problem. Nevertheless, the Benefit Integrity Project has now been established to check the validity of life awards of the highest rates of Disability Living Allowance (DLA). Although the DSS have described the Project

19 n 5 above, volume 2 261.

20 Now the fourth principle of the 1998 Act.

21 *Marcel* v *Commissioner of Police of the Metropolis* [1991] 1 All ER 845, 856.

22 This is particularly so given recent moves towards the setting up a National Intelligence Unit in the Contributions Agency: see Green Paper: *New Ambitions for Our Country – A New Contract for Welfare*, Cm 3805, (1998) 69.

23 HC Deb vol 286 col 66 25 November 1996.

24 K. Rowlingson *et al* (eds), *Social Security Fraud: The Role of Penalties* DSS Research Report No 64 (London: Stationery Office, 1997). The Report defines fraud as 'confirmed' if the fraud was admitted, or verified with an employer or evidence from the Home Office, irrespective of whether any further action led to prosecution: 23.

© The Modern Law Review Limited 1999

as a straightforward administrative exercise, it has its roots in the previous Government's drive against fraud, and has come in for severe criticism from many quarters.[25] Introduced without consultation with outside bodies, despite an undertaking by the Former Minister for State for Disabled People that consultations would be carried out, and prior to the new Government assuming office, despite the rules of conduct and good practice for General Elections dictating to the contrary, the Project has been criticised by the House of Commons Social Security Committee on two main points: claimants were wrongly losing their entitlement to benefit and the Project itself was being handled in an incompetent and insensitive manner.[26] The Committee's Report on DLA also drew attention to the fact that the Project concentrated only on the higher rates of benefit, and that there was no emphasis on investigating cases where claimants had not received their full entitlement.[27] In addition, while the initial review which prompted the introduction of the Benefit Integrity Project suggested that fraud was prevalent among DLA claims,[28] the level of fraud identified by the project was defined as 'minimal'.[29] The conclusion of the Social Security Committee was that the Project was seriously flawed conceptually and in practice.[30] The result is that the fears expressed prior to the implementation of the 1997 Act are now being realised.

Prosecution

Offences under the Social Security Administration Act 1992 were directed at claimants who made false claims on a regular basis. However, Housing Benefit is claimed for six to twelve months at a time so there are no repeat false statements and therefore no documentation on which to base a prosecution. As a result it was felt that prosecution in cases of alleged housing benefit fraud under the 1992 Act was 'practically impossible'.[31] The alternative is to prosecute under the Theft Act 1968. In 1996, the Court of Appeal ruled that a specimen offence could not be used to prove on-going criminal activity.[32] As a result, more individual instances of fraud have to be proved to demonstrate the extent of the on-going fraud. Consequently, prosecutions under the criminal law have become more difficult to conduct. The evidence before the Social Security Committee was that the difficulty with such prosecutions was exacerbated by the absence of specific expertise within Benefits Agency legal departments[33] and by the added burden such prosecutions placed on an authority's already limited resources.[34] This is compounded by the general reluctance of the police and Crown Prosecution Service to divert resources into prosecuting benefit fraud.[35] The central aim of the SSAF Act 1997 is to overcome the problems faced in prosecuting and punishing benefit fraud. Sections 13 and 14 amend section 111 of the Social Security Administration Act 1992 by

25 See Social Security Committee, Fourth Report: *Disability Living Allowance*, HC 641 (1998).
26 *ibid* para 54.
27 *ibid* para 56.
28 *ibid* para 48.
29 *ibid* para 57.
30 *ibid* para 71.
31 *ibid* volume 2, Q 214.
32 *R* v *Clark* [1996] 2 Cr App Rep 282. The implications for social security fraud were set out by Lord Mackay: 'unless the defendant agreed otherwise, the court would only sentence on the basis of the specimen offences, which might relate to benefit for a period of weeks out of a much longer period during which the fraud was perpetrated'. HL Deb vol 578 col 461 17 February 1997.
33 n 5 above, volume 2, Q 233.
34 *ibid* volume 1, para 88.
35 *ibid* volume 1, paras 85–87.

© The Modern Law Review Limited 1999

creating two new offences. Under section 13, if a person *dishonestly* makes a false statement or representation, furnishes false information, fails to notify a change of circumstances which is required by regulations to be reported, or causes another person to fail to notify such a change of circumstances, in order to obtain benefit he/she will have committed a criminal offence. Section 14 extends the existing offence of false representation for obtaining benefit under section 112 of the Social Security Administration Act 1992 so that an offence is committed if a person, *without reasonable excuse*, knowingly fails to notify a change of circumstances or knowingly allows another person to fail to notify a change of circumstances. Both offences are designed to tackle 'serious' social security fraud[36] but will apply equally to more minor frauds committed by individuals. However, numerous difficulties lie with these provisions.

Section 13 aims to take account of the 1996 Court of Appeal decision in *R* v *Clark*[37] and thus circumvent the problem of offenders being prosecuted for long running frauds on the basis of specimen evidence, allowing the court to take account of all payments received.[38] It may be argued that section 13 was a necessary addition to the 1997 Act, given the impracticalities of prosecuting each individual fraudulent representation. Nevertheless, the importance of legal safeguards in protecting innocent claimants should not be undermined. Given the complexity of benefits, the scope for innocent error is self-evident and, correspondingly, so too is the need to protect claimants from prosecution on this basis. It is also noteworthy that, in his evidence to the Social Security Committee, Neil Jackson of the DSS stated that the existing offences worked, despite difficulties in meeting the 'quite proper standards of proof required on a criminal charge', and pointed out that 'It should never be easy to hang a criminal offence around somebody's neck'.[39] With section 13, the prosecution must prove dishonesty on the part of the defendant before the offence is made out. That this is not the case with section 14 has been criticised by the Child Poverty Action Group (CPAG), who argue that the line between dishonesty and oversight or ignorance has been blurred.[40] Similarly, Alan Howarth has argued that the assumption behind section 14 is that all overpayments are the result of fraud.[41] In their Green Paper on welfare reform, the Government noted that claimants are sometimes unaware which changes they have to report, and how.[42] It must be realised that the main purpose of the 1997 Fraud Act is to detect and prosecute serious instances of fraud. Yet the individual claimant who inadvertently commits a minor fraud will also be covered and the possibility that a criminal offence may be 'easily hung around someone's neck' is an unsettling one.

Punishment

Once prosecutions are brought, success in achieving convictions is generally high,[43] not least because the fraud is generally documented – claimants will have signed a statement which can subsequently be proved to be false. However, the

36 HC Deb vol 286 col 66 25 November 1996.
37 n 32 above.
38 See HC Deb vol 286 cols 66–67 25 November 1996.
39 n 5 above, volume 2, Q 518.
40 *Welfare Rights Bulletin* 136, February 1997 8.
41 HC Deb vol 286 col 99 25 November 1996.
42 n 22 above 67.
43 98 per cent of prosecutions for benefit fraud result in a conviction: HC Deb vol 286 col 122 25 November 1996.

 © The Modern Law Review Limited 1999

likelihood is that sentences handed down by the court after conviction will be lenient. The evidence before the Social Security Committee was that a conviction in a magistrates' court usually resulted in either a fine or community service order,[44] and recent Divisional Court pronouncements by Staughton LJ suggest that fines imposed on those of limited means (such as benefit claimants) should be capable of being paid in a matter of weeks,[45] and so may not be commensurate with the scale of fraud committed. Such an approach has reinforced the perception that prosecution of offenders constitutes neither punishment nor deterrence. The response of the Social Security Committee was to recommend that appropriate sanctions for serious fraud be implemented, designed to have a deterrent effect.[46] However, under the SSAF Act, a new provision has been created to deal with the less serious frauds which never make it to court. Under section 15 of the Act, the Secretary of State or a local authority has the power to impose a financial penalty on a claimant as an alternative to prosecution. Where there are grounds for instituting proceedings against a person in relation to an overpayment the Secretary of State or local authority may give written notice, inviting that person to agree to pay a penalty. If the person agrees, then no proceedings will be instituted against him/her. The penalty is set at 30 per cent of the amount of the overpayment.[47] Agreement to pay may be withdrawn within 28 days. In addition, if, following the agreement to pay, it is decided 'on a review or appeal' that the overpayment is not recoverable, then any penalty already paid is to be returned.

Certainly it is desirable, for both claimants and taxpayers, for a case to be kept out of court where possible. According to former Secretary of State, Peter Lilley, there are some 100,000 cases a year where the DSS believe that there are grounds for prosecution but only 12,000 or so of those are taken to court. The anticipated effect of section 15 was that an additional 25,000 cases per year would face some form of prosecution.[48] Essentially, therefore, it was intended that section 15 would be invoked where the DSS believed that there were grounds for prosecution. The 1997 DSS Research Report on Social Security Fraud identified four levels of suspicion of fraud: where a claimant reports a change in their circumstances but not as soon as they should have; where there is a mild suspicion that fraud has been committed but no proof; where there is strong suspicion but no proof; and where the investigating officer is certain that fraud has been discovered but has insufficient evidence to prove it.[49] While it may be possible to make out a case in favour of the fourth level being an adequate ground on which to employ section 15, such an argument is likely to be tenuous and certainly could not be extended to levels one, two and three. The implication behind section 15 is that a prosecution *could* be brought, but that the levying of a penalty would be a more efficient and satisfactory means of dealing with the case. However, if the DSS present a case to the Secretary of State or a local authority on the basis that there is a strong suspicion or near certainty of fraud, the danger is that the Secretary of State or local authority will see the use of section 15 as expedient and therefore not explicitly require there to be supporting evidence sufficient to justify prosecution. This, in principle, would violate the purpose of section 15, in that suspicion of proof (at

44 n 5 above volume 1 para 90.

45 *R* v *Stockport Justices, ex parte Conlon; R* v *Newark and Southwell Justices, ex parte Keenaghan* [1997] 2 All ER 204.

46 n 5 above volume 1 para 90.

47 The penalty is in addition to the requirement that the claimant repays the overpayment: DSS Press Release, 26 May 1998.

48 HC Deb vol 290 cols 461–462 17 February 1997.

49 n 24 above 23.

© The Modern Law Review Limited 1999

whatever level) could become a substitute for the requirement that there are grounds for prosecution.

CPAG share these concerns and feel that the issue is whether the threat of prosecution is well founded to begin with.[50] Few people relish the prospect of court proceedings, and for many, alternatives to court, however unpalatable, will prove the more attractive option. The 'invitation' to pay a penalty rather than face prosecution for fraud is therefore likely to be accepted by a significant portion of claimants, who may not feel that they are in a position to argue. Where a suspicion of fraud is played off against a fear of court proceedings the result may be financially attractive for the Department[51] but it is one that smacks of manipulation. Allied with this is the problem identified by Alan Howarth in the House of Commons: 'I do not like to see such summary powers vested in officials. It is inevitable that over-zealous or bullying officials or agents will from time to time abuse them, particularly if they are incentivised to do so'.[52] Safeguards against such abuse do not appear to be built into the legislation. The only possibility claimants have of checking the evidence against them is to go to court, something which again many will be loath to do. While there is provision within section 15 for an 'appeal or review' it is not clear whether this can be instigated at the request of the claimant or if it is simply an internal procedure. The issue of incentives is also an important one. The indications are that local authorities must meet escalating fraud detection targets or risk subsidy loss.[53] The temptation may therefore be to meet targets through the use of section 15, even in instances when this is inappropriate.

Section 15 is intended to act as a deterrent for cases which are normally too trivial or time consuming to bring before the court. However, CPAG doubt that it will be truly effective, predicting that those with money will buy their way out, leaving those without to face prosecution.[54] Consistency in the application of section 15 therefore becomes an issue. If a landlord is discovered committing housing benefit fraud on a large scale, local authorities may opt for section 15 rather than instigate a formal prosecution, allowing them to meet their fraud detection targets with confidence. On this basis it is unlikely that the local authority will hesitate before applying section 15 to minor frauds, even where the evidence against the claimant is not solid. In this light, the dangers highlighted above can be seen to come into play. It is also clear that such developments are not in line with the Social Security Committee's recommendation that appropriate sanctions be implemented for *serious* fraud.

On a broader note, however, parallels can be drawn between section 15 and current thinking in criminal justice, where the concept of diversion has been prominent since the Criminal Justice Act 1991.[55] Dunbar and Langdon point out that both Labour and Conservatives subscribe to the general theory that lesser criminal offences should be diverted into a range of non-custodial disposals, including fines, and that pre-court disposals should be employed wherever possible.[56] This is in line with the thinking set out in the Green Paper on welfare

50 n 40 above 9.

51 Section 15 was expected to produce savings of £3.9 million per year: *ibid.*

52 HC Deb vol 286 col 100 25 November 1996.

53 See for example *Local Government Chronicle*, n 10 above. In their Green Paper the present Government have stated that local authorities will continue to be given financial incentives to improve their counter-fraud activity: n 22 above 68.

54 n 40 above 9.

55 See for example Uglow, *Criminal Justice* (London: Sweet and Maxwell, 1995) 259–267.

56 Ian Dunbar and Anthony Langdon, *Tough Justice: Sentencing and Penal Policies in the 1990s* (London: Blackstone Press Ltd, 1998).

© The Modern Law Review Limited 1999

reform, which accepts that fraud cases may be more appropriately dealt with away from the courts. Nevertheless, if alternatives to court and custody are to be an effective deterrent then they too must subscribe to the 'just deserts' principle which underpins the 1991 Act. The rhetoric of this has been adopted by the present Government in its Green Paper: 'Once a fraudulent payment or series of payments has been detected, the sanctions triggered should be fair and appropriate'.[57] While the principle of diversion can be seen to be both realistic and progressive, nevertheless the continued drive to be seen to be 'tough on crime' may come at the cost of many individual claimants. The deterrent effect of section 15 remains a disputed issue, as does the question of whether it is a 'fair and appropriate' sanction. It must be noted that, for an individual claimant, the amount of money involved in repaying the overpayment in addition to a thirty per cent penalty can represent a significant amount. This is particularly relevant if the fraud has been motivated by financial need rather than greed, something which section 15 does not take account of. Ultimately, therefore, the problems which section 15 creates look likely to outweigh its benefits.

Conclusion

The reforms introduced by the 1997 Act are obviously designed to save money and punish social security fraudsters, a policy endorsed by both sides of the party political divide. With losses from fraud at such high levels, obviously action is required. Nevertheless the problem of social security fraud needs to be put in perspective. The majority of fraud committed is as a result of organised crime yet there is a disproportionate emphasis in the 1997 Act on minor fraudsters. The view of the Social Security Committee on Housing Benefit Fraud and the evidence of the local authorities before it was that the prosecution of individual minor fraud achieves little. This view is not maintained by the 1997 Act. Data matching provisions, arguably created at the expense of certain basic human rights, assume not just that each claimant is *potentially* a fraudster but that he/she is *probably* so. Such an assumption is unfounded. Given the complexity of social security legislation, the provision of incorrect information or failure to report a change of circumstances is more likely to be due to innocent error than to fraud. Cuts in benefit administration can only contribute to this problem.[58] Pointing the finger at those claiming disability benefits in particular, where the facts do not bear out the original assumptions, contributes again to this anti-claimant culture.

The obligations of citizenship for social security claimants have been made very clear but little attention has been paid to any corresponding rights. Efforts to improve public awareness of fraud have not been matched with measures to improve the take-up of benefits, something which remains a problem.[59] Yet, as Ogus, Barendt and Wikeley point out, the indications are that underpayments exceed overpayments.[60] Cuts in benefit administration are likely to result in cuts in service standards, which may mean more claimants losing out on benefit entitlements through bad advice or lack of information. Claimants become further disadvantaged as restrictions on the back-dating of benefits come into

57 n 1 above 68.

58 See for example n 25 above 10–11; *Welfare Rights Bulletin* 137 April 1997 3.

59 For example, there is no national take-up campaign at the moment for Disability Living Allowance: see n 25 above, Minutes of Evidence, Q 271.

60 A. Ogus, E. Barendt and N. Wikeley, *The Law of Social Security* (London: Butterworths, 1995) 656.

© The Modern Law Review Limited 1999

play.[61] Dealing with the problem of fraud is entirely necessary and savings procured in this way are certainly preferable to those obtained by cutting benefit levels and entitlement. However, cuts in benefits also look likely, with benefit entitlement once again under threat.

More importantly, however, the 1997 Act does not address the underlying problems which can be said to lead to benefit fraud. As regards serious, organised fraud, the 1997 Act looks likely to improve matters for fraud investigators and prosecutors, although potentially at the expense of supposedly fundamental civil liberties. Individual claimant fraud, however, does not follow the same pattern as organised fraud and therefore demands a different approach. Dean and Taylor-Gooby suggest that a claimant's 'willingness to "fiddle" the system may primarily be a reflection of [their] adverse experiences or frustrations with that system'.[62] An examination of the structure of the benefits system reveals that there are perverse incentives within the system which enforce benefit dependency and encourage claimants to either commit minor fraud or face substantial financial insecurity. Despite recent improvements, such as working benefits, the system is not flexible enough to cope with the modern day realities of employment, or to accommodate rapid shifts in eligibility for means-tested benefits. As a result, in order to remain financially secure claimants are forced to withhold information about their changing circumstances, particularly where those changes are of a temporary nature. In addition, much of what is currently perceived as fraud is simply straightforward claimant error. Claiming social security has become an increasingly complex business, with the rules and regulations governing the administration of benefits growing steadily. Consequently, the potential for claimant error is endless. The pressure on claimants to get completed claim forms in on time can only add to the problem, and typically results in either delay for the claimant, or in the provision of inconsistent and inaccurate information thus giving rise to a suspicion of fraud. This note has emphasised the possibility of fraud arising from claimant confusion rather than dishonesty. The danger with the 1997 Act is that the distinction between the two will no longer be valid.

61 The Social Security (Miscellaneous Amendments) (No 2) Regulations 1997 SI 1997 No 793 cut the backdating provisions from twelve to three months.

62 Hartley Dean and Peter Taylor-Gooby, *Dependency Culture: The Explosion of a Myth* (London: Harvester Wheatsheaf, 1992) 119.

© The Modern Law Review Limited 1999

CASES

Profits from Wrongdoing: Private and Public Law Perspectives

Margaret Halliwell*

Public interest has been aroused recently when those who have committed crimes have been perceived to be making a profit from their crimes. The first case to attract publicity involved the payment of money to Mary Bell by Gitta Sereny. Further controversy followed in respect of money paid to the family of Louise Woodward and to the two nurses who were convicted of murder in Saudia Arabia. In the Bell case, the Attorney-General, in his role as protector of the public interest, investigated whether payment to Ms Bell could be blocked under a recent Court of Appeal case, *Attorney-General* v *Blake*,[1] barring George Blake, the double agent, from receiving £90,000 royalties from the sale of his memoirs.

The *Blake* case, which is the focus of this comment, develops the law in two important respects. First, it extends the availability of injunctive relief for the protection of the public interest where a breach of contract involves the commission of a criminal offence. Secondly, it extends the availability of restitutionary awards for breach of contract, thereby qualifying the principle that the purpose of damages awarded for breach of contract is simply to cover the losses of the innocent party (rather than to nullify any profit or savings made by the contract-breaker). Although the Court's remarks on this latter (restitutionary) point are strictly *obiter*, they foreshadow a major revision to the law of contract and invite careful analysis. Before looking at the restitutionary aspect of *Blake*, however, the circumstances of the case must be explained.

Attorney-General v *Blake*

The case arose because of the publication of the book, *No Other Choice*, by Jonathan Cape on 17 September 1990. The book is the autobiography of the convicted spy, George Blake, and contains information relating to his employment in the Secret Intelligence Service, his activities on behalf of the KGB, his trial, imprisonment and subsequent escape to Moscow. A sum of £90,000 remained payable to the defendant by the publishers. The Attorney-General was seeking to deprive the defendant of any financial benefit derived from publication of the book. In an earlier hearing, the Vice-Chancellor, Sir Richard Scott, had dismissed the argument that the copyright in the book and the profits derived therefrom belonged in equity to the Crown. The Attorney-General appealed against the finding that there was no private law remedy and, also, pursued an alternative remedy in public law. For reasons which will be explained, the public law claim succeeded and an injunction was granted restraining the defendant from receiving

* Faculty of Law, University of Manchester.

1 [1998] 1 All ER 833.

© The Modern Law Review Limited 1999 (MLR 62 2, March). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

any payment in connection with his autobiography. The Attorney-General had wished to keep a private law argument open for a higher court. By submitting the manuscript for publication, without having first obtained clearance, George Blake committed a clear breach of contract. He disclosed official information which he had contracted not to disclose. Under normal circumstances the Crown would only be entitled in contract law to nominal damages for this breach since no loss could be established. The argument was, however, that, in the circumstances of the case, the Crown might have a private law claim to restitutionary damages for breach of contract. Having obtained the public law remedy, it was abundantly clear that the Crown would not wish to appeal the case. George Blake had been fully represented throughout by *amici curiae*.[2] An *amicus curiae* has no standing to appeal and, for obvious reasons, the defendant had taken no part in the proceedings. Any observations on the restitutionary remedy were, therefore, unnecessary. The members of the Court of Appeal, however, considered the issue of such importance as to express their views even though *obiter* and being without the benefit of argument. These views are in line with the recent tendency to challenge the traditional compensatory remedy for breach of contract.

Challenges to the traditional view

The traditional rule, as formulated in the classic authority of *Robinson* v *Harman*,[3] provided problems in application where the measure of damages was capable of being addressed by way of either the 'cost of cure' or the 'diminution in value'. The prospect of the former measure provided an opportunity for defendants to obtain windfall profits whereas the latter measure might allow contract breakers to ignore obligations, particularly in building contracts with consumers. The House of Lords, in *Ruxley Electronics and Construction Ltd* v *Forsyth*,[4] was presented with precisely this type of building contract whereby a swimming pool had been built which did not match the contractual specification as to depth. The House of Lords remedied the problems by introducing, as well as those existing measures of damages, a hybrid measure to compensate for loss of amenity. This measure represented the 'consumer surplus' which is the measure to the plaintiff in excess of the market value of the breach. It has been suggested[5] that an appropriate measure of damages in such cases might be a restitutionary measure compelling the contract breaker to disgorge to the innocent party the amount of money representing the costs saved by the breach. The House of Lords , however, made it absolutely plain that the selection of the different measures would depend upon the principle of compensating plaintiffs. Compensatory damages in equity have also been developed to compensate, on a once and for all basis, for the effects of a future and continuing breach. Thus, in *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd*,[6] Brightman J, in circumstances where the

2 Lord Lester of Herne Hill QC and Mr Pushpinder Saini.

3 (1848) 1 Exch 850, 855 where Parke B said: 'The rule of the common law is that where a party sustains a loss by reason of a breach of contract he is, so far as money can do it, to be placed in the same situation with respect to damages as if the contract had been performed.'

4 [1996] AC 344: see J. Poole, 'Damages for Breach of Contract — Compensation and "Personal Preferences"' (1996) 59 MLR 272.

5 J. O'Sullivan, 'Loss and Gain at Greater Depth: The Implications of the *Ruxley* decision' in F. Rose (ed), *Failure of Contracts, Contractual, Restitutionary and Proprietary Consequences* (Oxford: Hart Publishing, 1997).

6 [1974] 2 All ER 321.

© The Modern Law Review Limited 1999

granting of an injunction would be inappropriate, awarded a sum of money which might have been demanded by the plaintiffs from the defendant as a *quid pro quo* for relaxing a restrictive covenant which the defendants had breached. There has been some doubt about the nature of such equitable damages. The best explanation, which insists that they are, also, truly compensatory, is given by Millett LJ in *Jaggard* v *Sawyer*.[7] This explanation will be considered below but is inter-related to the issue of whether, in any circumstances, English law will depart from the usual compensation principle and allow restitutionary damages where a contract breaker has profited from breach. It is in this context that the views of the members of the Court of Appeal in *Attorney-General* v *Blake* are important. First, they reflect the trend in the higher courts to adopt the principles of unjust enrichment[8] and to extend the circumstances when restitutionary remedies are available.[9] Second, they reflect the development of restitutionary remedies, as well as compensatory remedies, for a breach of contract.[10] Finally, they are indicative, in the area of unjust enrichment and restitution, of academic influence in rapid development of the law.[11] Before discussing the views as to the availability of a restitutionary remedy for breach of contract and the public law injunction, it is necessary to deal briefly with the appeal against the findings of Sir Richard Scott VC.

Fiduciary relationships and duties of confidence

The Court of Appeal accepted that the relationship between an employer and employees is of a fiduciary nature and that Blake, therefore, owed duties of trust and confidence to his employer. Such fiduciary duties, however, last only as long as the relationship which gives rise to them lasts. A former employee owes no such duties to a former employer. As to the fiduciary relationship of confidentiality, the Vice-Chancellor was correct to hold that the duty to respect confidence subsists only as long as the information remains confidential. The Crown's equitable claim against the defendant was, therefore, tenable on neither basis.

---

7 [1995] 1 WLR 269.

8 See Sir Donald Nicholls VC in *CTN Cash and Carry* v *Gallagher* [1994] 4 All ER 714, 720: 'In broad terms, in the end result the defendant may be said to have been unjustly enriched. Whether a new claim for restitution now, on the facts as they have since emerged, would succeed is not a matter I need pursue. I observe, as to that, only that the categories of unjust enrichment are not closed.'

For a full judicial endorsement of the principles of unjust enrichment, see the recent decision of the House of Lords in *Banque Financière de la Cité* v *Parc (Battersea) Ltd* [1998] 1 All ER 737.

9 In *Woolwich Building Society* v *IRC (No 2)* [1992] 3 All ER 737 a majority of the House of Lords established a principle of common law whereby money paid by a citizen to a public authority, in the form of taxes or other levies, and pursuant to an *ultra vires* demand by the authority, is *prima facie* recoverable by the citizen as of right.

10 Proponents of restitution advocate the adoption of The Law Commission's Report on *Pecuniary Restitution on Breach of Contract* No 121 (London: HMSO, 1983) which recommends that restitutionary remedies should be available to a part performer in breach. See A. Burrows *The Law of Restitution* (London: Butterworths, 1983) 281.

11 A notable illustration of this is the consideration given by members of the House of Lords in *Westdeutsche Landesbank Girozentrale* v *Council of the London Borough of Islington* [1996] 2 All ER 961 to W. Swadling, 'A New Role For Resulting Trusts' (1996) 16 LS, 110 and P. Birks, 'Restitution and Resulting Trusts' in D. Goldstein (ed), *Equity and Contemporary Legal Developments, Papers Presented at the First International Conference on Equity* (1992) 355.

© The Modern Law Review Limited 1999

The restitutionary remedy

The Court of Appeal accepted that the general rule is that damages for breach of contract are compensatory not restitutionary. The damages are measured by the loss to the plaintiff and not by the gain to the defendant. It was noted, however, that the elevation of the general rule into a fundamental principle which admits of no exceptions has been 'disrupted, attributed to inertia and has attracted widespread academic criticism'.[12]

The Court of Appeal considered judicial opinion to be divided with *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd*[13] providing an exception whereby the gain or the saving of expense made by the defendant is sometimes used as the measure of the plaintiff's loss. In contrast the Court of Appeal in *Surrey County Council* v *Bredero Homes*[14] refused to countenance the possibility of awarding restitutionary damages for breach of contract. The existing law can, however, be reconciled within the compensation rule. This is precisely what Millett LJ did, as mentioned earlier, in *Jaggard* v *Sawyer*.[15] In *Surrey County Council* v *Bredero Homes* the plaintiffs claimed damages for breach of a restrictive covenant against building more than seventy-two houses. They sought to measure the damages by reference to the additional profit which the defendant had made by building the extra houses. The distinction between that case and the *Wrotham Park* case is that, in the latter, damages were available in substitution for the award of an injunction under the jurisdiction originally conferred by s 2 of the Chancery Amendment Act 1858, commonly known as Lord Cairns' Act.[16] The Court of Appeal in the former case had no power to order an injunction. Millett LJ (in *Jaggard* v *Sawyer*) fully explained the significance of this distinction in the following terms:

> Both the Wrotham Park and Bredero Homes cases ... were concerned with a single past breach of covenant, so that the measure of damages at common law and under the Act was the same. Prima facie the measure of damages in either case for breach of covenant not to build a house on neighbouring land is the diminution in the value of the plaintiff's land occasioned by the breach. One element in the value of the plaintiff's land immediately before the breach is attributable to his ability to obtain an injunction to prevent the building. Clearly a defendant who wished to build would pay for the release of the covenant, but only so long as the court could still protect it by the grant of an injunction. The proviso is important. It is the ability to claim an injunction which gives the benefit of the covenant much of its value. If the plaintiff delays proceedings until it is no longer possible for him to obtain an injunction, he destroys his own bargaining power and devalues his right. The unavailability of the remedy of injunction at the same time deprives the court of jurisdiction to award damages under the Act and removes the basis for awarding substantial damages at common law.[17]

Clearly this explanation is based on an economic analysis of the bargaining power associated with a right to obtain injunctive relief for breach of a restrictive covenant. Millett LJ, moreover, emphasised that the award made by Brightman J in the *Wrotham Park* case was indeed compensatory, not restitutionary. Brightman J did not award the plaintiff the profit which the defendant had made by the breach. He awarded the amount which he judged the plaintiff might have obtained as the price of giving its consent. The amount of the profit which the defendant expected

12 n 1 above, 844.
13 n 6 above.
14 [1993] 1 WLR 1361.
15 n 7 above.
16 The jurisdiction is now to be found in s 50 of the Supreme Court Act 1981
17 n 7 above, 212.

 © The Modern Law Review Limited 1999

to make was a relevant factor in the assessment, but that was all. The analysis by Sir Thomas Bingham MR in *Jaggard* v *Sawyer* was in very similar terms. He could not accept the fact that Brightman J's assessment of damages in *Wrotham Park* was based on other than compensatory principles. Brightman J's attention to the profits earned by the defendants was paid, not in order to strip the defendants of their unjust gains, but because of the obvious relationship between the profits earned by the defendants and the sum which the defendants would reasonably have been willing to pay to secure release from the covenant. If this analysis of existing law by the Court of Appeal in *Jaggard* v *Sawyer* is correct, which it is in my opinion, the novelty of the judiciary recommending, in *Attorney-General* v *Blake*, restitutionary damages for a breach of contract, becomes apparent. Millett LJ, in the *Jaggard* case, showed no antipathy towards the award of restitutionary damages for a breach of contract. He merely stated, in the context of *Surrey County Council* v *Bredero Homes* that 'Unless they were able to recover damages in accordance with restitutionary principles, neither at common law nor in equity could the plaintiffs recover more than nominal damages'.[18] Given Lord Woolf's views as to the role of judges in developing the law and Millett LJ's judicial and extra-judicial activities in pursuing the development of restitution, it is, therefore, not surprising that in their view, and in the view of Mummery LJ, a private law remedy was available to enable the Crown to claim a restitutionary remedy against George Blake for his breach of contract. Their views as to the general circumstances when restitutionary remedies will be available for breach of contract, although strictly *obiter*, will, undoubtedly, have considerable influence on the future development of the law.

The circumstances when restitutionary remedies will be available for a breach of contract

According to the Court of Appeal in *Attorney-General* v *Blake*, the basis on which restitutionary damages should be available for breach of contract will be exceptional. In particular, the defendant's moral culpability will not alone form the basis of such an award. Also, the fact that the breach of contract is deliberate and cynical will not, by itself, be a good ground for departing from the normal basis on which damages are awarded. More importantly, at least from the perspective of contract lawyers who advocate the importance of economically efficient breach,[19] the mere fact that the defendant's breach of his contract with the plaintiff has enabled him/her to enter into a more profitable contract with someone else should also not be sufficient. The views of their Lordships, therefore, satisfy more traditional, if not out-moded, classical contract doctrines of freedom and sanctity of contract whereby parties, in general, are free both to make and to break contracts.

In the course of concluding that two exceptions will apply, one for skimped performance, the other a bespoke *Blake* exception, the Court of Appeal considered the recommendations of the Law Commission in its Consultation Paper No. 132 *Aggravated Exemplary and Restitutionary Damages*.[20] In enunciating the two particular exceptions, the Court of Appeal acknowledged that they do not precisely

18 n 7 above, 212.

19 Whereby economic efficiency and the best allocation of resources is reached by maintaining legal rules which optimise the circumstances when a contract breaker can choose to pay damages to the innocent party in order to make greater profits in an alternative transaction elsewhere.

20 (London: HMSO, 1993) para 7.7 p 159.

© The Modern Law Review Limited 1999

coincide with the provisional conclusion of the Law Commission's Consultation Paper. They concluded, however, that the second exception falls squarely within the two conditions identified in para 7.20 of the Consultation Paper. These conditions are that: (a) there was deliberate wrongdoing which could have been restrained by injunction and; (b) in addition, the gains made by the defendant are 'attributable to the interest infringed' in the sense that they are referable to the breach of an obligation which the defendant had specifically contracted not to breach.

The two exceptions

The case of skimped performance

This is where the defendant fails to provide the full extent of the services which he has contracted to provide and for which he has charged the plaintiff. The paradigm is the Louisiana case of *City of New Orleans* v *Fireman's Charitable Association*.[21] In that case, the defendant contracted with the plaintiff to provide a fire fighting service and was paid the full contract price. After the expiry of the contract the plaintiff discovered that the defendant had not provided the stipulated number of firemen or horses or the promised length of hosepipe. In other words, the defendant had saved itself substantial expense by the breach. The plaintiff was unable to recover more than nominal damages as the defendant had not failed to put out any fires in consequence of the breach. According to the views expressed by the Court of Appeal, in such a case, English law will allow substantial restitutionary damages measured by the amount which the defendant has saved by the breach. This exception appears to embrace the suggestion[22] that in cases such as *Ruxley Electronics and Construction Ltd* v *Forsyth*[23] a restitutionary remedy will, in future, be available to allow for disgorgement of a sum equivalent to the costs saved. Clearly, such a remedy will depend upon absolute proof that costs were saved by the breach. There is little doubt that the builders in *Ruxley* made no appreciable saving from the breach. Where, however, such savings can be proved, it appears that a plaintiff will have four possible remedies: the cost of reinstatement; the difference in cost to the builder of the actual work done and the work specified; the diminution in value of the work due to the breach; or, as in *Ruxley*, a measure to compensate for loss of amenity.[24] Such a battery of remedies may well be of benefit in consumer contracts where there has been no loss in market value as a result of the breach and where, therefore, compensation is necessarily hard to quantify. The difficulty in this extension of remedies is that consumers will, understandably, attempt to insist on the most advantageous award. In the case of skimped performance the measure of damage, if proven, is reasonably clear. This will not, however, prevent a plaintiff in consumer contracts from arguing a number of alternative measures, particularly when damages for 'cost of cure' or 'loss of amenity' may well yield a greater sum in such contracts. Similarly, defendants will remain keen to argue for the 'diminution in value' measure of damages.

The addition of a restitutionary remedy poses no real problem to normal contract principles as there will have been no conscious allocation of risk. Consumers will,

21 (1891) 9 So 486.

22 n 5 above.

23 n 4 above.

24 For the first three see Lord Cohen in *East Ham Corporation* v *Bernard Sunley & Sons* [1966] AC 406 cited by O'Sullivan n 5 above, 20.

© The Modern Law Review Limited 1999

ordinarily, expect the contractual specifications to be met. To avoid disruption of general contract principles there will, however, need to be a deliberate effort to observe the economic arguments underpinning the normal measure of damages for breach in commercial contracts. O'Sullivan contends[25] that a restitutionary remedy by way of disgorgement of profit should not apply where the plaintiff's interest in performance is merely to make profit or substitute performance can be readily bought on a genuine market. However the distinction is made, it is essential that in their zeal to develop the law of restitution, the judges should not encroach too far into well-established contract principles and this leads to a consideration of the second exception enunciated in *Attorney-General* v *Blake*.

Where the defendant has obtained his profit by doing the very thing which he contracted not to do

This exception covered the facts of the present case exactly. George Blake promised not to disclose official information and he did so for profit. He earned the profits by doing the very thing which he had promised not to do. The restitutionary remedy in such circumstances requires the disgorgement of such profits made from the breach of contract.

The Court of Appeal concluded that the second exception falls squarely within the two conditions identified in para 7.20 of the Law Commission's Consultation Paper since there was both deliberate wrongdoing which could have been restrained by injunction and, in addition, the gains made by the defendant were 'attributable to the interest infringed' in the sense that they were referable to the disclosure of official information which the defendant had contracted not to disclose. Finally, the members of the Court of Appeal expressed the joint view that only time will tell as to whether the distinctions made are tenable. They concluded, nonetheless, that the general rule that damages for breach of contract are compensatory can safely be maintained without denying the availability of restitutionary damages in exceptional circumstances. This may well be true. Certainly, the second exception to the general rule is rooted in the particular facts of the case at hand. This exception, however, countenances the very development of the law which the members of the Court of Appeal in the *Surrey County Council* case had sought so hard to resist. In entering into the restrictive covenant the developers promised not to build more houses than the contract provided for. The deliberate breach of this covenant led to the profits being earned. The developers earned the profits by doing the very thing which they had promised not to do. The potential for restraint by injunction did, of course, exist. Even stronger arguments arise for the disgorgement of profits in cases similar to the *Wrotham Park* case since the possibility of restraint by injunction was actual, rather than merely potential. The Court of Appeal in *Attorney-General* v *Blake* mentioned a common feature between the two exceptions which may be helpful, in the future, in addressing this point. This feature is that, in both exceptions, compensatory damages are an inadequate remedy if regard is paid to the objects which the plaintiffs sought to achieve by the contract. This overriding point is clearly evident in the circumstances of skimped performance in consumer contracts but, as has been discussed, appears merely to add to the existing remedies available and not to replace them in any meaningful way. The best remedy available will remain the compensatory one of 'cost of cure' and, in my opinion, the availability of this remedy should not be undermined by the new restitutionary remedy.

25 n 5 above.

© The Modern Law Review Limited 1999

The overriding point does not seem to inhibit restitutionary claims in the context of breach of restrictive covenants since, in almost all cases, the very objective of the covenant is to prevent the breach which, ultimately, occurs. It is difficult to assess, particularly in the light of the aforementioned lack of antipathy on the part of Millett LJ to the introduction of restitutionary remedies, the degree of innovation which the views of the Court intended to achieve. The explanation which Millett LJ gave as to the distinction between *Wrotham Park Estate Co Ltd* v *Parkside Homes Ltd*[26] and *Surrey County Council* v *Brodero Homes*[27] is based on an economic analysis of the bargaining power associated with a right to obtain injunctive relief for breach of a restrictive covenant. As explained earlier Brightman J, in the former case, awarded the amount which he judged the plaintiff might have obtained as the price of giving its consent. He did not award the plaintiff the profit which the defendant had made by the breach. The measure of damages was, rightly, compensatory. The council in the latter case had, however, failed to capitalise on the inherent economic bargaining value of the potential for injunctive relief. Such bargaining value did not exist at all in the case of *Attorney-General* v *Blake*. The Crown had no interest in granting permission for Blake to disclose the information. Had Blake sought permission for publication it would not, at any price, have been granted. The potential for injunctive relief did not have the economic value described by Millett LJ in *Jaggard* v *Sawyer*. The Crown, therefore, had no bargaining rights on which to found a compensatory claim for equitable damages. In my opinion, this feature of *Blake* provides a sufficient ground for departing from normal contractual principles and allowing the restitutionary claim. Where market economics prevail, no such claim should succeed. In other words, both the *Wrotham Park* case and the *Surrey County Council* case were rightly decided. The contractual compensatory principle should apply in such circumstances to give rise, where appropriate, to compensation in equity based on a sum representing the price which might have been paid for the release of the covenant. Arguments for the introduction of a restitutionary remedy to disgorge profits should be resisted on the grounds of economic efficiency. Whilst economic efficiency can never usually form the criterion for development of the law, it should at least play some part in the context of purely commercial contracts. To introduce restitutionary remedies in such contracts would be introducing, via the back door, a general requirement of good faith into contract law. Such a development may be desirable by way of a substantive change in contract law but not by way of initiating new remedies.

It must be emphasised that the whole tenor of the decision in *Attorney-General* v *Blake* is admittedly cautious. Both cynical breach and profit making are excluded as giving rise to a restitutionary claim. The Court of Appeal acknowledged that Steyn LJ, in *Surrey County Council* v *Bredero Homes*, gave persuasive reasons as to why the circumstances when restitutionary remedies should be available for breach of contract should remain exceptional. In a case decided after *Attorney-General* v *Blake* was reported, *Gafford* v *Graham*,[28] Nourse LJ delivered the judgment of the Court of Appeal which wholly endorsed the views taken by both Sir Thomas Bingham MR and Millett LJ in *Jaggard* v *Sawyer*. It is to be hoped that this analysis continues to be adopted when awarding damages in lieu of an injunction. As stated, however, the two exceptions may well admit to more circumstances than was intended.

26 n 6 above.
27 Text to note 16.
28 *The Times* 1 May 1998 (Lexis Transcript)

© The Modern Law Review Limited 1999

From the perspective of private lawyers, activism and innovation amongst judicial proponents of the law of restitution is commonplace. From the perspective of public lawyers, the role of the present judiciary in active law making is often obscured by the technicality of restitution law. It is this issue, in the context of the public law claim in *Attorney-General* v *Blake*, which must now be addressed.

The public law claim

The Attorney-General was construed as seeking relief in his historic role as guardian of the public interest. In *Gouriet & Ors* v *The Union of Post Office Worker & Ors*[29] all the members of the House of Lords accepted that the Attorney-General is entitled to invoke the aid of the civil courts in support of the criminal law and that a court has the power to grant injunctive relief in an appropriate case in support of the criminal law. The House of Lords had, however, stressed that, because civil courts should be reluctant to become involved in criminal proceedings, the Attorney-General's powers should be limited to cases where an offence is frequently repeated in disregard of a usually inadequate penalty, or, to cases of emergency. In the view of the Court of Appeal in *Attorney-General* v *Blake*, the powers are not necessarily so limited. The principle enunciated by Lord Diplock in *Erwen Warnink* v *J. Townsend & Sons Hull* was particularly helpful:

> ... where over a period of years there can be discerned a steady trend in legislation which reflects the view of successive Parliaments as to what the public interest demands in a particular field of law, development of the common law in that part of the same field which has been left to it ought to proceed upon a parallel rather than a diverging course.[30]

In this situation, the legislative development to uphold the policy that a criminal does not retain profit directly derived from the commission of a crime is demonstrated by confiscation orders in the Criminal Justice Act 1988 (Part IV) as amended by the Proceeds of Crime Act 1995. That statutory procedure could not be applied in the circumstances of George Blake's serious criminal offence, the breach of the Official Secrets Act. This provided good reasons for the Attorney-General to uphold public policy by obtaining the injunction sought. Safeguards against the abuse of the Attorney-General's powers would be provided by a number of factors: the court will carefully scrutinise the application to ensure that it is in the interests of justice; an injunction is unlikely to be granted unless it is providing carefully targeted support to the enforcement of substantive legislation enacted by Parliament; and, the court would be very mindful of avoiding a result which unjustifiably interfered with free speech. On the latter point, the Court of Appeal emphasised that the grant of the injunction was readily justified under the European Convention on Human Rights as being a necessary and proportionate measure adopted in the interests of national security.

Conclusion

The views of the Court of Appeal typify the willingness of judges to develop the law of restitution and to indulge in creative law making at the same time. As in many restitution cases, academic opinion was thoroughly canvassed and the Court

29 [1978] AC 435.
30 [1979] AC 731, 743.

© The Modern Law Review Limited 1999

of Appeal cited ten national and international academic sources.[31] Another typical aspect of the rapid development of the law of restitution in recent years has been the willingness of the judiciary to consider reform of the existing law. The relationship between the legislature and the judiciary is often at issue in such creative law making. Although the Court of Appeal sought to follow the policy of legislation in developing the common law, judges are not always so circumspect. Lord Goff, in *Woolwich Building Society* v *IRC (No 2)*[32] was particularly unimpressed with the objection that reformulation of the law would upset the boundary between the legitimate development of the law by the judges and the role of legislation. He did not think the boundary to be clearly drawn. More conservative views prevailed in another restitution case, *Westdeutsche Landesbank Girozentrale* v *Council of the London Borough of Islington*.[33] In that case the majority of the House of Lords agreed with Lord Browne-Wilkinson that the introduction of compound interest in a common law action would be a usurpation of the function of Parliament. Lord Goff and Lord Woolf, in the minority, were not so persuaded. Further judicial development to prevent profits being made as a result of crime may prove unnecessary as the current government, following the recent publicity about such cases, has promised to strengthen the existing legislation. It is not inconceivable, however, that the law of restitution may develop further in this area.[34]

It remains to be seen to what extent restitutionary damages will be awarded in future cases to disgorge profits gained by breach of contract. The members of the Court of Appeal in *Blake* were clearly anxious to delineate the two exceptions from the general compensatory rule as carefully as possible but the potential for further erosion of the rule in the future is far-reaching. There is also a problem as to precisely what profits have to be disgorged. In the case of skimped performance the measure, if proven, is reasonably clear. The second exception is much less clear. The Court of Appeal appeared to make the assumption that Blake should make no profit from the publication of his book and yet a great deal of the information in the book was not relevant to the breach of his contract with the Crown. Obviously, but for his breach of contract, his autobiography might have had little, if any, value. Nevertheless, the Court seemed to proceed on the basis that, once the exception is established, all profits from the breach should be disgorged. If this is indicative of a general rule, then unless this exception is construed extremely narrowly, freedom to break contracts, pay damages and contract elsewhere, will be consigned to history. It is essential, therefore, that future judicial interpretation of the decision should be undertaken with extreme care.

31 n 1 above, 844.
32 n 9 above.
33 n 10 above.
34 See the arguments put by D. Friedmann, 'Restitution for Wrongs: The Basis of Liability and the Measure of Recovery' in W. Cornish (ed), *Restitution: Past Present & Future* (Oxford: Hart Publishing 1998) 113.

© The Modern Law Review Limited 1999

Contributory Negligence applied to Economic Loss: *Platform Home Loans* and *Fancy and Jackson*

Tony Dugdale*

Section 1 of the 1945 Law Reform (Contributory Negligence) Act provides: 'Where any person suffers damage as the result partly of his own [the claimant's] fault and partly of the fault of any other person [the defendant], a claim in respect of that damage shall not be defeated by reason of the fault of the person suffering the damage, but the damages recoverable shall be reduced to such extent as the court thinks just and equitable having regard to the claimant's share in the responsibility for the damage'. For almost 50 years the 1945 Act was limited to cases of physical damage and its application was largely uncontroversial. In the 1990s recognition that concurrent tort and contractual actions lie and that the contributory negligence regime applies to both, has led to the 1945 Act being applied to cases of pure economic loss. But the process has not been as straightforward as in the case of physical damage. It tends to be clear whether a plaintiff has contributed to his own physical injury. The causation and nature of economic loss is more complex with the result that the plaintiff's conduct may be regarded as contributing to some but not all aspects of the loss. Again, whilst it is clearly negligent for a plaintiff to take a risk with his own physical safety, it may be perfectly sensible for him to run the risk of economic loss in the hope of making a profit. These problems have surfaced in the extensive litigation between lenders and professionals arising from the property recession of the early 1990s. Two recent decisions, *Platform Home Loans Ltd* v *Oyston Shipways Ltd*[1] and *Bristol and West BS* v *Fancy & Jackson*[2] focus on the meaning of contribution and negligence in this context, and raise broader questions about the role and impact of contributory negligence in the context of commercial relationships.

The property recession resulted in lenders, who had made loans on the security of borrowers' property, claiming losses suffered upon default from valuers responsible for negligently overvaluing the security at the time when the loan was made and/or solicitors responsible for negligently failing to warn of problems concerning the security or the borrower. Initially the courts took the view that the lender could recover all its loss on the transaction from the negligent professional where it would not have entered that transaction had it been given proper advice. In *Banque Bruxelles*[3] Lord Hoffmann considered this transaction loss measure was not 'fair or reasonable'. Rather, in the case of the negligent valuer responsibility should be limited to the consequence of the information being wrong, namely, that the lender made an advance on less security than he had been led to believe by the overvaluation. That lack of security, the amount of the overvaluation, was the

* School of Law, Keele University.

1 [1998] 4 All ER 252, CA.

2 [1997] 4 All ER 582.

3 *Banque Bruxelles Lambert SA* v *Eagle Star Insurance Co Ltd* [1997] 1 AC 191; also reported as *South Australia Asset Management Corp* v *York Montague Ltd* [1996] 3 All ER 365, HL. Noted in (1998) 61 MLR 68.

© The Modern Law Review Limited 1999 (MLR 62:2, March) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

attributable loss and the lender was entitled to recover his actual loss on the transaction only to the 'cap' of the missing security. Valuers and solicitors have sought to further limit the lender's recovery by claiming that their lending policies such as that of making non-status loans (where there is no investigation of the borrower's ability to repay) both contribute to the damage and constitute fault so as to justify a reduction of damages under the 1945 Act. In *Platform* the Court of Appeal considered whether such policies could be contributory whilst *Fancy & Jackson* concerned fault and the reduction of damages.

Contributory conduct

In *Platform* the plaintiff had lent a borrower just over £1 million secured by a mortgage on a house which the defendant had valued at £1.5 million. The loan was on a non-status basis. The borrower defaulted and the lender resold the property for £435,000. The evidence suggested that the true value of the house at the date of the loan was £1 million. Following *Banque Bruxelles*, the lender's maximum claim was for the amount of the overvaluation, that is, £500,000. The valuer argued that the lender's claim should be further reduced on account of his contributory negligence in making a non-status loan. Jacob J held that it was imprudent to make a non-status loan of over £1 million where that constituted 70 per cent of the value of the security and reduced the lender's damages by 20 per cent.[4] The lender appealed on the ground that, even if imprudent, its non-status lending policy did not contribute to the damage suffered by the plaintiff and, hence, there could be no reduction under the 1945 legislation. The argument turned on the construction of the word 'damage' in the 1945 Act. The lender argued that it should be construed narrowly as referring to just the loss recoverable against the defendant and that, applying this definition, its policy did not contribute to that loss as it had 'nothing to do with the overvaluation'.[5] The valuer countered that 'damage' referred to all the loss sustained by the plaintiff and not just that element recoverable against the defendant. On this basis, the lender's policy was contributory because it had increased the chance of default and loss on the transaction. The lender's narrower construction had been accepted in a number of first instance decisions,[6] essentially by analogy with the reasoning in *Banque Bruxelles*. There it had been held that a valuer was liable for only the losses which fell within the scope of his duty and not for all those which were the factual result of his breach. Parity of treatment might suggest that the lender should be liable to have its damages reduced only where the recoverable loss fell within the scope of the risk created by its conduct and not just because the recoverable loss was a factual result of that conduct.[7] If the reason for

4 It was agreed that five per cent of this reduction was attributable to the lender's admitted fault in not questioning the borrower's failure to give the purchase price of the property as required by the mortgage application form. The lender accepted that its conduct was contributory in this respect.

5 The application of the argument can be questioned as the decision to lend at a 70 per cent loan to value (LTV) ratio combined with the overvaluation to result in the inadequacy of the security. See further Evans, 'Contributory Negligence by Lenders' (1998) 14 PN 43.

6 *Interallianz Finanz AG* v *Independent Insurance Co* (5 June 1997, unreported); *Coventry Building Society* v *William Martin & Partners* [1997] 48 EGLR 159; *Britannia Building Society* v *Hallas* (4 July 1997, unreported); *Bristol & West Building Society* v *Fancy & Jackson* [1997] 4 All ER 582; *Britannia Building Society* v *White & Co* (13 August 1997, unreported).

7 In *Causation in the Law* (Oxford: OUP, 2nd ed, 1985) 208, H L A. Hart and T. Honoré explain that there are two views as to the meaning of 'contributing cause' in the context of contributory negligence. Conduct may be regarded as contributing if it was a necessary element for the production of harm. Alternatively, conduct may be regarded as contributing if the eventual harm was within the

© The Modern Law Review Limited 1999

holding a non-status loan to be negligent is the increased risk of the borrower defaulting and loss being sustained even if the security had been correctly valued, then arguably the loss flowing from the inadequacy of the overvalued security is not within the risk. Thomas J put the point in *Interallianz Finanz AG* v *Independent Insurance Co Ltd*[8]: '[c]onsistent with the decision in *South Australia* ... the only relevant negligence that can be taken into account in reducing ... damages is negligence that is referable to the recoverable damages. ... as the valuer's duty does not extend to advice on the overall prudence of entering into the loan transaction (which would include the credit review of the borrower and assessment of the market risk), negligence in relation to those matters is ... irrelevant'.

Giving the only judgment of the Court of Appeal in *Platform*, Morritt LJ rejected the narrow construction, holding that the earlier decisions had wrongly 'equated the "damage" to which the section refers for the purpose of prescribing the conditions for its operation with the "damages" recoverable from the valuers which, if those conditions are satisfied, the section empowers the court to apportion'.[9] The parity of treatment argument was not considered. Rather, the lender put its case in terms of the injustice of allowing a 'double deduction for the fault of the lender'; the first being the deduction of losses going beyond the amount of the overvaluation under the *Banque Bruxelles* principle and the second being the further reduction of the capped recovery for contributory negligence. Morritt LJ dismissed this argument on the ground that it would lead to the bizarre conclusion that the apportionment legislation would not apply where the loss had been capped under *Banque Bruxelles* but would apply where the loss had fallen within the cap which would usually be the case when the property market was stable. He concluded that Parliament could not have intended the Act to be construed in this way. This must be right. The possibility of a double deduction is relevant to the justice of any reduction under the Act and not to the construction of the Act. However, this reasoning does not meet the parity of treatment point which rests not on the possibility of a double deduction but on the allocation of responsibility and risk between the parties.

Having dismissed the lender's justice argument, Morritt LJ based his broad construction of damage on the authority of *Froom* v *Butcher*.[10] There the defendant's negligent driving caused an accident and the plaintiff's failure to wear a seat belt resulted in head injuries in addition to the hand injury which would have resulted even if a belt had been worn. Citing Lord Denning's view that the damage was 'caused in part by the bad driving, and in part by the failure ... to wear a seat belt', Morritt LJ concluded that 'the "damage" ... in that case included damage for which the claimant was exclusively responsible and was properly taken into account as it was part of the overall loss sustained by the plaintiff'.[11] Applying this reasoning to the lending context, the term 'damage' would include loss stemming exclusively from the lender's conduct. Thus, it would include losses in excess of the amount of the overvaluation which had been caused by the lender's failure to discover that the borrower's covenant to repay was worthless. But the analogy with

risk, that is, it was harm 'the likelihood of which is among the reasons for holding the [conduct] to be negligent'. They conclude their survey of the case law, all of which related to physical injury, by stating that it was difficult to determine which approach had been followed as the line between the two was 'very slender' *ibid* 211–12. Arguably, the line is clearer in economic loss cases and the risk principle easier to apply.

8 [1997] EGCS 91.

9 [1998] 4 All ER 252, 262, CA.

10 [1976] 1 QB 287, CA.

11 [1998] 4 All ER 252, 261, CA.

© The Modern Law Review Limited 1999

Froom does not quite hold. The claimant in *Froom* was not 'exclusively responsible' for the head injury for his failure to wear the belt did not break the chain of causation and, hence, the defendant remained liable for that injury.[12] A better road accident analogy with *Platform* would be one where a plaintiff was responsible for risky driving leading to an accident but a defendant garage was responsible for inadequately securing the seat belt so that it was unable to protect the plaintiff from head injury. The defendant would be liable for the head injury but not loss unrelated to the wearing of the belt, for example, damage to the car. The plaintiff would be exlusively responsible for the latter damage. It seems likely that the plaintiff's conduct in causing the accident would be regarded as contributing to the head injury. To that extent the analogy supports the broad construction in *Platform*.

Perhaps the strongest argument in favour of the broad construction is one which features not in the judgment of Morritt LJ but in that of Judge Lloyd QC in *Mount Credit Corporation Ltd* v *Abbey National Estate Agency Ltd*,[13] delivered the day before that in *Platform*. This is an argument based on practicality. Judge Lloyd described attempts to distinguish between a lender's conduct contributing to the transaction loss and that contributing to the narrower recoverable loss, as being symptomatic of 'Thomist' logic and inappropriate in the context of the apportionment legislation. The point can be illustrated by reference to the facts of *Platform* where the interrelationship of valuation, loan to value ratio (LTV) and creditworthiness was apparent from the lending policy in question which scaled the LTV from a maximum 92 per cent on loans of £150,000 where the creditworthiness of the borrower was investigated to one of 70 per cent LTV on non-status loans of over £500,000. It is difficult to separate one aspect of the lending policy from another. Arguably, it is the policy as a whole which contributes to the lack of security compounded by the overvaluation. The suggestion that elements could be separated with the LTV element being regarded as contributory to the inadequacy of the security whilst the non-status element is non-contributory may be difficult to sustain. Whether it follows from this example that all attempts to draw distinctions are inappropriate, is questionable.[14] Nevertheless, there is an attraction in treating all the claimant's conduct as relevant to contributory negligence and sweeping up the fine distinctions in the net of a just and equitable approach to the reduction of damages. As Morritt LJ suggested in *Platform*, the court will 'no doubt pay regard to the fact that the lender will be solely responsible [for the loss exceeding the overvaluation cap]'.[15] The consequence of this approach will be to shift attention from the concept of contribution and the meaning of the term 'damage' to the concept of fairness between the parties and the meaning of the term 'fault'.

12 A true case of exclusive responsibility would be one where a claimant suffered cuts in a road accident caused by his negligence and then suffered additional damage as a result of the negligence of a doctor. The doctor's conduct would break the chain of causation and he would be exclusively responsible for the additional damage with the claimant being exclusively responsible for the accident injuries (see *Robinson* v *Post Office* [1974] 2 All ER 737). There is no authority to suggest that such a plaintiff's damages against the doctor could be reduced to take account of his conduct in causing the original accident.

13 Unreported, 18 December 1997.

14 In *National Home Loans Corp plc* v *Giffen Couch & Archer (a firm)* [1997] 3 All ER 808, 815, CA, Phillips LJ, albeit in the context of the scope of a solicitor's retainer, considered that there was a fundamental distinction 'between points going to security in the event of default . . and points going to the worth of the personal covenant'.

15 [1998] 4 All ER 252, 262, CA

© The Modern Law Review Limited 1999

Fault

At first instance in *Platform*, Jacob J found the lending policy to be at fault because it failed to consistently recognise that the reliability of domestic valuations decreased as the supposed value of the property rose. The policy recognised this at lower valuations and provided for a scaled lowering of the loan to value ratio as the level of valuation increased. But the progression stopped at requiring a 70 per cent LTV at a £50,000 loan level. No extra precautions, such as a lower LTV or a status inquiry, were taken in relation to the £1 million lent in the case and that was the fault. The lender was condemned by its own policy. Whilst this particular decision on fault was uncontroversial and not appealed, the approach to assessing fault endorsed by Jacob J does raise a more difficult issue. He cited with approval the dictum of Sir John Vinelott in *Birmingham Midshires Mortgage Services Ltd* v *David Parry*[16] that fault was not to be tested on the basis of the common practice of lenders. Sir John recognised that if the question had been whether lenders were in breach of a duty of care to others, common practice would normally be conclusive of the issue. The lender would not be liable because he could not be 'expected to be wiser than his fellows'.[17] But he argued that a different standard applied when contributory negligence was in issue: 'evidence of the way in which other businesses are conducted is not a reliable guide to the question whether a business was conducted prudently – that is whether those conducting the business took reasonable care to protect themselves against the risk of injury or loss. There may be good commercial reasons which lead those engaged in a business enterprise to take risks'.[18]

The underlying issue is one of how prudence is to be judged. In *Midshires*, Sir John Vinelott suggested that a non-status loan at 75 per cent LTV was imprudent because the 25 per cent cushion would barely cover the typical costs of repossession and resale and provide no cover for any market fall. However, if the non-status loan policy is considered as a whole, the risk of loss on a minority of loans may well be more than balanced by the prospect of profits on a majority of loans. It is a question of swings or roundabouts. Viewed as a swing, an individual non-status loan may go down and be regarded as imprudent. As part of the roundabout of a loan portfolio, it may be regarded as part of a prudent commercial policy. As Wright J said in an earlier case, commercial lenders are 'by definition ... in business to take a degree of risk' and they guard against the risk by taking security.[19] From the lenders' perspective the policy could be seen to be based on a prudent level of risk. However, from the perspective of the professional acting for the lender in an individual transaction, the policy increases his exposure to liability. It makes it more likely that the borrower will default and the security will prove inadequate. The application of the prudence test in *Midshires* suggests that the lender must act prudently in relation to the potential liability of the professional. It must not impose an unreasonable risk on the professional.

Such thinking seems to underlie the treatment of lending secured by a mortgage indemnity guarantee (MIG) in *Bristol and West BS* v *Fancy & Jackson*.[20] There the lender had lent 98 per cent of the value of the property but, as was then common practice amongst lenders, insisted that the element of the loan in excess of 75 per cent

16 [1996] PNLR 494. The defendant solicitor was held not liable and no decision on contributory negligence was required

17 [1996] PNLR 494, 518.

18 [1996] PNLR 494, 518.

19 *H I T Finance* v *Lewis & Tucker* [1993] EGLR 231, 236. Wright J rejected the notion that the lender was contributory negligent in failing to investigate the credit status of the borrower.

20 [1997] 4 All ER 582.

© The Modern Law Review Limited 1999

of the value of the property be covered by an MIG policy. The defendant solicitor had negligently failed to warn the lender of matters relating to the borrower's status and applying the *Banque Bruxelles* test, was liable for the lender's entire loss on the transaction.[21] Chadwick J held that the lender's decision to lend 98 per cent of the value was 'negligent or imprudent whether or not the party has insured against the consequential loss'.[22] The MIG policy was to be ignored in assessing the prudence of the loan just as a plaintiff's personal accident insurance policy would be ignored in assessing the negligence of his conduct which contributed to the accident causing his injuries. The lender's damages were reduced by that portion of the loss attributable to lending in excess of 75 per cent of the property value.

The personal accident analogy is not perfect. Chadwick J argued that where a person 'undertakes some hazardous activity and is injured partly as a result of his own negligence and partly as a result of the negligence of another, it could be no answer to [a contributory negligence] apportionment ... that the injured party had effected insurance against the very injury which occurred'.[23] That is clearly correct. A personal accident policy does not prevent the physical damage to the speeding motorist. It is just as imprudent to speed with a policy as without one. An MIG policy does prevent the only kind of loss a lender can suffer: financial loss. It is surely as prudent to lend money on the combined security of property and an MIG policy as on the strength of property or a guarantee alone. The MIG policy is part of the security. But the situations are analogous in another way. The negligent motorist injuring another who is covered by an accident policy, will not benefit from the proceeds of that policy. Such proceeds are not deducted from the damages recoverable. In the same way, the MIG payment is not deducted from the lender's recoverable loss. This is because the lender is bound to repay the MIG insurer from any damages recovered.[24] So, although the MIG benefits the lender by providing additional security, it provides no benefit to the negligent professional. From his perspective the liability risk involved in a 98 per cent loan backed by an MIG is the same as that involved in a 98 per cent loan without such backing. By making a 98 per cent loan on a property with an MIG the lender may be exposing the professional to a greater risk of liability than it would making a 75 per cent loan without an MIG.

If the test for 'fault' is whether the plaintiff has exposed the defendant to a greater risk of liability, then we are back to asking questions about the 'contribution' of the plaintiff's conduct to the loss recoverable from the defendant. Although *Platform* holds that the lender's conduct does not have to contribute to the recoverable loss to be relevant, it may still have to contribute to the valuer's exposure to liability to be found blameworthy. The point can be illustrated by reference to the MIG issue in *Fancy & Jackson*. Exceptionally, in that case the relevant defendant was liable for the whole loss on the loan including the element covered by the MIG.[25] However, in the normal case where the valuer's or solicitor's negligence went to the value of the security rather than the status of the

21 The consequence of failing to warn as to status was that a loan was made to a borrower unable to repay with the result that the lender had to call on the security. All the losses suffered as a result of calling on the security were attributable to the solicitor.

22 [1997] 4 All ER 582, 625.

23 *ibid.*

24 *Bristol & West Building Society* v *May May & Merrimans (No 2)* [1997] 3 All ER 206.

25 This was because the defendant solicitor, Steggles Palmer, was in breach of a fiduciary duty to the lender to report that it was acting for the vendor as well as the borrower/purchaser and that the vendor was the employer of the purchaser. Had the lender known that, it would not have proceeded with the transaction whatever the valuation of the security. Hence, all the losses resulting from having to call on the security were recoverable.

© The Modern Law Review Limited 1999

borrower, the recoverable loss would be limited to the element of the loan secured on the property.[26] The approach taken by Chadwick J would allow for no contributory negligence reduction in such a case for in *Fancy & Jackson* he reduced the damages by that portion of the loss attributable to lending in excess of 75 per cent of the property value. In the normal case, that portion would not be recoverable anyway and hence, no reduction would fall to be made. An alternative way of expressing this outcome is to say that where conduct, such as lending beyond 75 per cent LTV on the security of the MIG, imposes no extra risk on the professional then it cannot be regarded as being blameworthy.

The 'just and equitable' reduction

Two problems arise in this context. The first is that the equity of any reduction will have to be considered in the context of the recoverable loss. If we take the example of a loan at £80,000 on a valuation of £100,000, then following the reasoning of Vinelott J in *Birmingham Midshires*, the LTV of 80 per cent might be regarded as imprudent as it might be anticipated that on default the lender would face costs exceeding its £20,000 cushion. For example, where the cost of resale was £10,000 and the loss of interest between default and resale was £20,000, the total costs of £30,000 would exceed the cushion by £10,000. Suppose that on default the property is resold for just £80,000 and that the lender suffers such a £30,000 loss on the transaction: If, as in *Fancy & Jackson*, a professional was liable for that full loss on the transaction then it might be perfectly just to reduce the damages by deducting the loss attributable to the high LTV ratio. If it would have been prudent to loan £70,000 rather than £80,000, the approach of Chadwick J would suggest that £10,000 should be deducted from the recoverable loss. But the equity is different where the professional is only liable to the cap of the overvaluation. Suppose that a valuer has valued the property at £100,000 when its true valuation was £80,000. Again, the lender suffers a £30,000 loss on the transaction but here its recovery is limited to the £20,000 overvaluation cap. The lender's £10,000 shortfall has been caused by its failure to allow a sufficient cushion. To apply a further £10,000 deduction or, indeed, a percentage reduction for contributory negligence would seem to unfairly penalise the lender, imposing a double deduction on the lender in respect of its failure to allow a sufficient cushion. Like the 'contribution' argument, the 'double-deduction argument' was rejected in *Platform* in the context of the construction of the Act; like the 'contribution' argument, it will return to influence the application of the Act.

A second problem stems from the court's discretion in determining the fair level of reduction. Judges have suggested a range from 75 per cent,[27] through 30 per cent,[28] 25 per cent[29] to the 20 per cent applied in *Platform*.[30] Some degree of predictability is necessary to assist the settlement process. One answer would be to adopt the approach taken by Lord Denning in the seat belt case of *Froom* v *Butcher*

26 In *Bristol & West Building Society* v *May May & Merrimans (No 2)* [1997] 3 All ER 206 at 214, Chadwick J stated that in the normal case, '[p]rima facie, the part of the loss attributable to the negligent advice is that part of the loss which results from the amount of the normal advance having been fixed at a figure higher than it would have been but for the negligent advice'.

27 *First National Commercial Bank* v *Andrew S Taylor Ltd* [1997] 1 PNLR 37 (obiter).

28 The reduction applied by Phillips J in *Banque Bruxelles Lambert SA* v *Eagle Star Insurance Co Ltd* [1995] 2 All ER 769. See also *Midland Bank Plc* v *Douglas Allen* (unreported, July 1997).

29 *Bristol & West Building Society* v *Kramer & Co* [1995] NPC 14; *Cavendish Funding Ltd* v *Henry Spencer & Sons Ltd* [1998] PNLR 122, CA.

30 And also in *Nyckeln Finance Co Ltd* v *Edward Symmons & Partners* [1996] PNLR 245.

© The Modern Law Review Limited 1999

and fix the percentage, in that case at 25 per cent where the belt would have prevented all injury and 15 per cent where there would still have been some injury. Such a tariff is reasonably straightforward to apply where there is just one form of faulty conduct in issue and expert evidence to determine the injuries which would have been avoided. It will be more difficult to develop in the lending context where there are a number of forms of faulty or risky conduct and the consequences of such conduct do not fall into two neat categories. The alternative may be to follow the approach of Chadwick J and deduct particular items of loss which can be attributed to the conduct of the lender.

The broader picture

It has been suggested that one interpretation of the lending cases is that the courts have considered the lender to be at fault in exposing the professional to greater risk of liability. The difficulty is that the lender has not hitherto been regarded as owing any duty of care in relation to its lending policies to the professionals on whom it relies. Limiting the lender's recovery to reflect the imposition of risk on the professional might be thought tantamount to imposing such a duty. Wright in his essay, 'The Standards of Care' in the collection *Philosophical Foundations of Tort Law*,[31] argues that in the context of contributory negligence, the court should consider only the risks to the plaintiff's own well-being and not those to others because a requirement to 'weigh others' interests equally with your own ... would completely undermine individual autonomy and freedom'. This view would suggest that exposing the professional to greater risk should only amount to fault if the lender has accepted a responsibility to the professional in that respect. In other words, where the circumstances are such that a court could imply a term that such care be taken in relation to the lending policy.[32]

An alternative approach is suggested by Simons' essay on 'Contributory Negligence' in the same collection.[33] Simons suggests that there are two possible approaches to the definition of fault in the context of contributory negligence. It can be regarded as meaning 'conduct that is deficient relative to a standard of reasonable care'.[34] It is implicit in such an approach that society prefers that the plaintiff should have acted otherwise. Alternatively, it could encompass risky conduct in which the plaintiff engages for the benefit likely to result from the risk. Society might not wish to deter such activity but might consider that the plaintiff should 'pay for the risky behaviour to the extent of suffering a reduction in recovery against the defendant. In other words, the defendant should not be required to "subsidize" the plaintiff's risky behaviour'.[35] Simons notes that 'current doctrine employs similar criteria for contributory negligence and injurer negligence'[36] but puts this down to familiarity

31 D. Owen (ed) (Oxford: Clarendon Press, 1995) 249–275, particularly 254–259, 269–270.

32 Where the relationship is purely tortious in nature, the same result could be achieved by imposing a reciprocal duty of care on the lender or limiting the scope of the valuer's duty to the context of lending policies embodying such a reciprocal element

33 n 3 above, 461–485.

34 *ibid* 464.

35 *ibid* 466. Simons terms this a strict liability or strict responsibility rule but the essence is the notion of risk. Later in the essay at 483, he suggests that the criteria for applying a strict responsibility rule could be extrapolated from the rationale for defendant's strict liability developed by Fletcher, namely, that one person should not impose on the other a 'non-reciprocal risk' or obtain from the other a 'non-reciprocal benefit'. Applied to the lender scenario, it can be said the lender's risky non-status lending both generates a non-reciprocal benefit for the lender and imposes a non-reciprocal risk on the professional.

36 *ibid* 478.

© The Modern Law Review Limited 1999

with the negligence label rather than incisive analysis. By contrast, he suggests that courts have not explicitly articulated 'a strict responsibility' rationale and notes that risky conduct has been seen as relevant only to the narrow defence of 'assumption of risk'. However, the context in which the courts have been hitherto considering the notion of contributory fault has been that of physical damage, a context in which it is natural for the law to discourage deficient activity as defined by the negligence standard. The context of pure economic loss is entirely different. Society has no general interest in discouraging particular forms of conduct. Indeed, the notion of a standard for business conduct conflicts with the nature of competition which drives business to innovate rather than standardise. But society does have an interest in a fair allocation of losses between those concerned. One interpretation of the approach taken by the courts in the lending cases is that it is based on risk rather than fault and has the objective of securing a fair allocation of losses.

Whatever the theoretical basis of contributory negligence, the recent case law has confirmed it as a major piece in the chess game being played between the legal teams representing lenders and professionals; a piece on the professionals' side of the board and a piece whose rules of movement are not yet settled. Lenders may respond in one of two ways to this problem. The first is exclusionary: they might try to remove the piece from the board. This could be achieved by redrafting the standard instructions to professionals so as to contain strict duties. No reduction for contributory negligence can be made where the defendant is in breach of a strict duty.[37] It could be achieved more directly by an exclusion in the professional's retainer of any apportionment for contributory negligence. Such an exclusion would not normally be subject to the Unfair Contract Terms Act 1977 as the controls in that Act are limited to clauses which directly or indirectly restrict or exclude liability. It does not apply to clauses which remove a limitation on liability which would otherwise apply.[38] Lenders may have the commercial power[39] to impose such strict duties or exclusions on professionals but, leaving aside the damage to the relationships, there is always the possibility that the tactic might be thwarted by the judges. They have been reluctant to accept that strict instructions to report particular matters do any more than amplify the general duty of care.[40] Very clear wording will be required to establish a strict duty. The exclusion clause route may also be thwarted if the judges are prepared to imply that the lender owes the professional a duty in relation to lending polices, for then the effect of the clause will be to exclude a liability and the Unfair Contract Terms Act will apply.

The alternative approach would be inclusionary. Lenders could accept the reciprocity of their relationship with professionals and seek to control the unpredictability of the contributory negligence regime by expressly allocating responsibilities and risks in the retainer. This points to the wider role that contributory negligence may play in the context of economic loss. Its potential application may persuade parties to set their own standards of reciprocity. It may lead to an adjustment of the parties' behaviour rather than simply an adjustment of compensation and that is to be welcomed.

37 See for example *Barclays Bank plc* v *Fairclough Building Ltd* [1995] 1 All ER 289, where it was held that damages could not be reduced to take account of the plaintiff's contributory negligence where the defendant was in breach of a strict duty to provide 'workmanship of the best of its kind'.

38 Unless 'liability' is construed as encompassing a liability to have damages reduced under the Contributory Negligence provisions.

39 Lenders control the market for professional services by the use of 'panels' of approved firms working to standardised retainers. For comment on the liability implications, see A. Sprince 'The liability of solicitors to lenders on borrower default' (1998) 14 PN 3.

40 *Mortgage Express Ltd* v *Newman & Co* [1996] PNLR 603.

© The Modern Law Review Limited 1999

'Hearsay-fiddles' in the House of Lords

John Hartshorne* and Andrew L.-T. Choo**

The hearsay rule provides that a statement made otherwise than by a person giving oral evidence in court is inadmissible as evidence of the truth of its facts. A hearsay statement becomes admissible if one of the common law or statutory exceptions to the rule applies.[1] In 1964 the House of Lords held in *Myers* v *DPP*[2] that the list of common law exceptions to the hearsay rule was closed; any further exceptions would have to be created by Parliament. Confessions of a defendant represented a well established common law exception to the hearsay rule.[3] Arguably this exception only permitted the *prosecution* to adduce the confession; any exception permitting a defendant to adduce a confession of a co-defendant, in circumstances where that confession had either been excluded for the prosecution or not relied upon by them, was one unknown to English law.[4] In the recent decision of the House of Lords in *R* v *Myers*,[5] however, their Lordships allowed the confession of a co-defendant not relied upon by the prosecution to be adduced by a defendant as evidence of the truth of its facts. It will be argued that this represents the creation of a new common law exception to the hearsay rule by means of a 'hearsay-fiddle'.[6] The aim of the House of Lords was clear; to avoid the mechanical application of the hearsay rule in a way which would shield the truth. This is to be applauded. Yet by resorting to a 'hearsay-fiddle' rather than a more direct acknowledgment that the law of hearsay was being recast in a less restrictive mould, the House of Lords has left uncertainty about the operation of this new exception. We aim in this note to illustrate this uncertainty and to evaluate the soundness of their Lordships' reasoning. In doing so, we highlight the dangers of resorting to 'fiddles' as opposed to changing the law in an open and coherent manner.

The facts and background

Myers and Quartey were jointly charged with the murder by stabbing of a taxi driver. The defendants ran 'cut-throat' defences, each alleging that the other was

* Faculty of Law, University of Leicester.
** Department of Law, Brunel University.

1 See A. L.-T. Choo, *Evidence: Text and Materials* (London: Longman, 1998) 321–358

2 [1965] AC 1001.

3 This common law exception has now been placed on a statutory footing by section 76(1) of the Police and Criminal Evidence Act 1984 which provides: 'In any proceedings a confession made by an accused person may be given in evidence against him in so far as it is relevant to any matter in issue in the proceedings and is not excluded by the court in pursuance of this section.'

4 Opinion is divided on the question of whether section 76(1) permits a defendant as well as the prosecution to adduce a confession. See M. Zander, *The Police and Criminal Evidence Act 1984* (London: Sweet & Maxwell, 3rd ed, 1995) 219.

5 [1997] 3 WLR 552. For discussion of this decision elsewhere see M. Hirst, 'Confessions as Proof of Innocence' (1998) 57 CLJ 146.

6 An expression borrowed from D. J. Birch, 'Hearsay-Logic and Hearsay-Fiddles: *Blastland* Revisited' in P. Smith (ed), *Criminal Law: Essays in Honour of J C. Smith* (London: Butterworths, 1987).

© The Modern Law Review Limited 1999 (MLR 62 2, March) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

responsible for the stabbing. Significantly, Myers twice confessed whilst she was in police custody that she was the one who had 'cut' the taxi driver.[7] The prosecution elected not to rely upon the confessions which, having been made in circumstances involving several breaches of the Codes of Practice issued under the Police and Criminal Evidence Act 1984,[8] would have been liable to be excluded under section 78 of the Act.[9]

Defence counsel for Quartey submitted that he was entitled to adduce the confessions in Quartey's defence, even though the prosecution had elected not to rely upon them. The introduction of the confessions for Quartey would be probative of the fact that it was Myers and not Quartey who had carried out the stabbing. In considering this submission, the trial judge referred to two decisions of the Court of Appeal. The first was *R* v *Beckford and Daley*,[10] where it was held that the repetition by police officers of a co-defendant's confession which had been excluded for the prosecution pursuant to section 78 of the Police and Criminal Evidence Act 1984 would constitute inadmissible hearsay for a defendant. In the second case, *R* v *Campbell and Williams*,[11] it was held that a defendant could rely upon an informal admission obtained from a co-defendant. *R* v *Beckford and Daley* was distinguished on the ground that the confession in that case had been excluded for the prosecution; by contrast *Campbell and Williams* concerned an admission which would have been admitted for the prosecution had they sought to rely upon it. To prevent a defendant from relying upon such a confession would defeat common sense.

The trial judge in *R* v *Myers* took the view that a statement against interest by one party should be provable against that party by another so long as both remained parties to the action. Evidence of the confessions was adduced and Myers was convicted of murder and Quartey of manslaughter. In the Court of Appeal the trial judge's ruling was upheld.[12] The Court of Appeal granted Myers leave to appeal to the House of Lords, certifying that the following point of law of general public importance was involved in its decision:

> In a joint trial of two defendants, A and B, is an out of court confession by A which exculpates B but which is ruled, or is conceded to be, inadmissible as evidence for the Crown nevertheless admissible at the instigation of B in support of B's defence, or does such a confession in all circumstances offend the rule against hearsay?

The decision in the House of Lords

In logic one would have expected their Lordships to be faced with three possible options in reaching their decision. They could have held either (1) that the

7 The first confession was made during a conversation with a police constable at the police station. The second was made to officers who were escorting her in a vehicle taking her to the magistrates' court.

8 No contemporaneous records were made of the confessions and Myers was not invited to sign any written record of the confessions (Code C paragraph 11) nor was she under caution when she made the confessions (Code C paragraph 10).

9 Section 78(1) provides: 'In any proceedings the court may refuse to allow evidence on which the prosecution proposes to rely to be given if it appears to the court that, having regard to all the circumstances, including the circumstances in which the evidence was obtained, the admission of the evidence would have such an adverse effect on the fairness of the proceedings that the court ought not to admit it.'

10 [1991] Crim LR 833.

11 [1993] Crim LR 448.

12 [1996] 2 Cr App R 335. On the Court of Appeal's decision see A. L.-T. Choo, 'The Hearsay Rule and Confessions Relied upon by the Defence' (1997) 1(3) E & P 158.

© The Modern Law Review Limited 1999

adduction by a defendant of a confession which could not be adduced by the prosecution was permitted as an established exception to the hearsay rule; or (2) that no such exception existed and none could be created in accordance with the prohibition laid down in *Myers* v *DPP*; or (3) that no such exception existed but one ought to be created and the straitjacket of *Myers* v *DPP* discarded. Lord Slynn, with whom Lord Hutton and Lord Steyn agreed, held that Quartey had been correctly permitted to adduce evidence of Myers' confessions, and dismissed the appeal. In reaching this decision his Lordship chose an option which in reality was not available; that a new exception to the hearsay rule should be created to allow a defendant to adduce a confession of a co-defendant, but that this would not involve a departure from the straitjacket of *Myers* v *DPP*.

Lord Slynn acknowledged that there was a direct conflict between the Court of Appeal authorities of *R* v *Beckford and Daley* and *R* v *Campbell and Williams*, but thought that the latter should be followed.[13] His Lordship noted that a defendant is permitted to cross-examine a co-defendant upon a confession for the purposes of showing inconsistency and hence damaging credibility.[14] If the making of the confession is denied it can be proved against the co-defendant pursuant to section 4 of the Criminal Procedure Act 1865. A defendant remains entitled to cross-examine a co-defendant on a confession and prove its making even if the confession is excluded for the prosecution.[15]

Lord Slynn then took a giant intellectual leap by stating that a defendant should also be allowed to introduce a co-defendant's confession in evidence *for its truth*, by way of cross-examination either of the co-defendant or of witnesses to whom the confession was made. This is subject, however, to the very important caveat that such cross-examination would not be allowed if the confession was obtained in a manner which would make it inadmissible for the prosecution pursuant to section 76(2) of the Police and Criminal Evidence Act 1984.[16] As there was no suggestion that Myers' confessions had been obtained under any of the circumstances referred to in section 76(2), Lord Slynn held that the confessions had been rightly admitted on Quartey's behalf. Disappointingly, Lord Slynn concluded his speech by refusing to answer the certified point of law on the basis that it went much wider than the facts of the case at hand.[17]

Lord Hope, with whom Lord Mustill agreed, gave his own reasons for dismissing the appeal. It is not easy to decide which option his Lordship selected from the three referred to earlier. It is perhaps most accurate to say that Lord Hope believed he was endorsing an existing exception to the hearsay rule.[18] His Lordship was heavily influenced by discussion to be found in two Scottish cases, *McLay* v *HM Advocate*[19] and *Perrie* v *HM Advocate*,[20] to the effect that a defendant ought to

13 Whilst Lord Slynn expressly set the question which of the conflicting authorities ought to be followed, his Lordship did not expressly answer it. It is implicit from his Lordship's judgment that *R* v *Campbell and Williams* ought to be followed. The avoidance of an express answer is part of the 'fiddle' inherent in this decision

14 Pursuant to *R* v *Rowson* [1986] QB 174 and *Lui Mei Lin* v *The Queen* [1989] AC 288.

15 *ibid.*

16 [1997] 3 WLR 552, 563. Section 76(2) provides that a confession must be excluded if it was or may have been obtained by oppression of the person who made it (section 76(2)(a)) or in consequence of anything said or done which was likely, in the circumstances existing at the time, to render unreliable any confession which might be made by him in consequence thereof (section 76(2)(b)).

17 [1997] 3 WLR 552, 564.

18 Lord Hope approved the decision in *R* v *Campbell and Williams*; to that extent his Lordship was endorsing an 'existing' exception.

19 1994 JC 159.

20 1991 JC 27.

© The Modern Law Review Limited 1999

be permitted to adduce the confession of a co-defendant where this would be of relevance to his or her defence.[21] Lord Hope found this Scottish reasoning reflected in the judgment of Hobhouse J in *R* v *Campbell and Williams*, and concluded accordingly that 'the reasoning on the issue of hearsay which is contained in *Reg.* v *Campbell and Williams* ... is to be preferred to that in *Reg.* v *Beckford and Daley*.'[22] Having effectively decided that the adduction by a defendant of a co-defendant's confession would not offend any rule against hearsay,[23] his Lordship added the qualification that a confession would be inadmissible as defence evidence if it had been obtained in any of the circumstances referred to in section 76(2) of the Police and Criminal Evidence Act 1984. The confession would be inadmissible not because section 76(2) extended to confessions upon which the defence proposed to rely, but because if the confession had been obtained by oppression or in circumstances conducive to unreliability, it would be 'worthless for all purposes'.[24] Like Lord Slynn, Lord Hope refused to answer the certified point of law.[25]

A hearsay fiddle?

It is argued that Lord Slynn created a new exception to the hearsay rule in the face of *Myers* v *DPP*, and that Lord Hope endorsed an existing exception to the hearsay rule which in reality did not exist at all. This argument presupposes, of course, that prior to *R* v *Myers*, there was no exception to the hearsay rule allowing a defendant to adduce a confession of a co-defendant as evidence of the truth of its facts. Support for this position is drawn from the Court of Appeal in *R* v *Rowson*[26] and the Privy Council in *Lui Mei Lin* v *The Queen*,[27] where the question was addressed of whether it was permissible for a defendant to cross-examine a co-defendant on a previous inconsistent confession which had been excluded for the prosecution. In both decisions it was held that it was, but that the jury should be warned in clear and careful terms that they must not rely upon the confession as evidence of the truth of its facts, but only as evidence relevant to the co-defendant's credibility.[28] If prior to the decision in *R* v *Myers* there *was* an exception to the hearsay rule which permitted a defendant to adduce a confession of a co-defendant as evidence of the truth of its facts, one would have thought that such an exception would have been expressly referred to. It was not. Indeed if such an exception did exist there would have been no need to have stressed the importance of warning the jury that the confession was only evidence going to credibility.

21 It must be noted that Scottish law is not precluded from creating new common law exceptions to the hearsay rule in the same way that English law is pursuant to *Myers* v *DPP*. Lord Hope noted, however, that in the decision of *McLay* v *HM Advocate* the Lord Justice-Clerk, Lord Ross, said that any alteration to the hearsay rule as applied in criminal cases would have to be made by Parliament. Lord Hope concluded that 'on this point I think that the law of the two countries can be taken to be virtually indistinguishable': [1997] 3 WLR 552, 569. Lord Slynn, having referred to *Perrie* v *HM Advocate*, said 'in the law of Scotland the categories of hearsay evidence are not closed': [1997] 3 WLR 552, 562.

22 [1997] 3 WLR 552, 570.

23 Like Lord Slynn, Lord Hope expressly avoided saying whether a new exception to the hearsay rule was being created or whether an established exception was being endorsed.

24 [1997] 3 WLR 552, 571.

25 *ibid*, 572.

26 [1986] QB 174.

27 [1989] AC 288.

28 In *Lui Mei Lin* v *The Queen*, Lord Roskill said of the confession that 'its only relevance is to test the credibility of the evidence which the maker of the statement has given against his co-accused': [1989] AC 288, 298.

© The Modern Law Review Limited 1999

Further support for this position is drawn from the fact that in *R* v *Beckford and Daley*, *R* v *Campbell and Williams* and *R* v *Myers* itself, there is not to be found a single reference to any English authority which recognises such an exception. The 'hearsay fiddle' of *R* v *Myers* is that a new exception to the hearsay rule was created with a judicial sleight of hand that managed to conceal the fact that this was being done and that the principle in *Myers* v *DPP* was being outflanked. Reluctance to acknowledge this openly probably explains why Lord Slynn and Lord Hope were so coy about providing an answer to the certified point of law.

Unresolved issues

Three important points arise from the decision in *R* v *Myers* which were not addressed adequately by the House of Lords. First, the House of Lords blurs the distinction between evidence admissible for its relevance to credibility and evidence admissible for the truth of its facts. This distinction is crucial in the law relating to the admissibility of previous consistent statements and previous inconsistent statements. Lord Slynn's reasoning in *R* v *Myers* was effectively that if Myers' confessions were admissible before the jury as previous inconsistent statements, then they might as well be admissible as evidence of their facts. In other words, if the jury were to hear this evidence at all, they might as well hear it for all purposes rather than just one. If this reasoning is taken to its logical conclusion, however, *all* evidence that is relevant to the defendant's case and which has hitherto been admissible only as evidence going to credibility will potentially be admissible now as evidence going to the truth of its facts. It is disappointing that this important issue was not explored further by the House of Lords.

Secondly, the House of Lords' decision creates an anomaly between the admissibility of confessions of co-defendants and those who are not a party in the proceedings. Both Lord Slynn and Lord Hope expressly approved the rule in *R* v *Blastland*[29] that a confession made by a person who is not a party in the proceedings is inadmissible hearsay. It now follows that if A and B are tried jointly, A's confession may be adduced by B by way of exception to the hearsay rule. If A and B are tried individually, however, A's confession is inadmissible hearsay for B. It is difficult to see how this distinction can be justified, especially if in a joint trial a co-defendant elects not to testify. There can be no logical distinction between the situation where the person who made the confession is present in court but does not testify, and the situation where the person who made the confession is not a party in the proceedings at all. Lord Hope was of the view that different rules should apply to co-defendant confessions because a co-defendant cannot be compelled to testify, and a defendant may therefore be deprived of the opportunity of adducing evidence relevant to his or her defence.[30] This rests on the assumption that a non-party *will* be compellable to testify and will incriminate him- or herself in the witness box. Assuming that a non-party could be compelled, however, if he or she denied making the confession, the decision in *R* v *Blastland* would prevent the defendant from adducing evidence of it from other sources. A justification based upon non-compellability is thus unconvincing. It might be argued, of course, that confessions made by a non-party, being easy to

29 [1986] AC 41.
30 [1997] 3 WLR 552, 568.

© The Modern Law Review Limited 1999

fabricate and therefore potentially unreliable, are not in any event worth admitting in evidence. This overlooks the fact, however, that a confession of a co-defendant may be just as unreliable, especially in those areas where the decision in *R* v *Myers* will have its greatest impact: where the confession is excluded for the prosecution because of breaches of the Codes of Practice. Given that the House of Lords had the courage to create a new exception catering for co-defendant confessions, it is frustrating that the rule in *Blastland* was accepted so unquestioningly.

Thirdly, the decision sits uneasily with the traditional principle that the hearsay rule applies with equal rigour to the defence as it does to the prosecution. This has led to the exclusion of highly relevant defence evidence in celebrated decisions such as *Sparks* v *R*[31] and *R* v *Harry*.[32] In the former, the defendant was charged with indecent assault on a young girl and wished to call the girl's mother to testify that the girl had said that 'it was a coloured boy' who had committed the crime. The defendant was white. The court ruled that this evidence was inadmissible hearsay. Unjust though this decision may have been, it was the logical outcome of the application of the hearsay rule. There is no logical reason why a defendant such as Quartey should not be subject to the full rigour of the hearsay rule but a defendant such as Sparks should continue to be. In both cases the defendants were seeking to adduce evidence of statements made out of court for their truth. Again, it is frustrating that the House of Lords had nothing to say about the continued validity of such a distinction.

The test of admissibility

The new exception permits a defendant to adduce evidence of an out-of-court confession of a co-defendant, provided that the confession is relevant and that it was not made in circumstances which would infringe the prohibitions contained in section 76(2) of the Police and Criminal Evidence Act 1984. If the confession does infringe the prohibitions in section 76(2) then it will be inadmissible; such evidence is considered to be 'worthless for all purposes'.[33] It follows that a confession which infringes section 76(2) of the Police and Criminal Evidence Act 1984 is inadmissible for a defendant, whereas one which infringes section 78 of the same statute is presumptively admissible. The soundness of this distinction between section 76(2) and section 78 is questionable. First, trial judges are not always consistent in their use of section 76(2) and section 78. There are many cases where trial judges use section 78 to exclude a confession which ought more appropriately to have been excluded under section 76(2).[34] Secondly, the distinction almost presupposes that a confession excluded under section 76(2) is so unreliable as to be regarded as worthless for all purposes, whereas a confession excluded under section 78 is reliable enough to be used by the defence. In practice a confession excluded under section 78 may be just as unreliable as one excluded under section 76(2).

It should be pointed out that Lord Hope acknowledged that a confession excluded under section 78 may not always be available to the defence, but declared that 'this

31 [1964] AC 964.

32 (1986) 86 Cr App R 105.

33 [1997] 3 WLR 552 per Lord Hope at 571. This was an ingenious way of avoiding having to decide the vexed question of whether section 76(1) expressly applies to confessions upon which the defence propose to rely. Both Lord Slynn and Lord Hope chose to avoid directly answering this particular question. See note 4 above.

34 See M. Zander's commentary on section 78 in *The Police and Criminal Evidence Act 1984*, n 4 above.

© The Modern Law Review Limited 1999

is a point on which we did not hear argument, and it is not necessary to reach a view upon it in order to reach a decision in this case'.[35] One could hazard a guess that Lord Hope envisaged depriving the defence of this evidence on the ground that such a confession would be 'worthless'. It should not be forgotten, however, that as the law currently stands, there is meant to be no discretionary basis upon which admissible defence evidence can be excluded.[36] Exclusion of evidence on the grounds of 'worthlessness' is therefore a new concept to the law of evidence. Having chosen to introduce this concept, Lord Hope ought to have explained it properly. As a result of his Lordship's failure to do so, many questions now remain unanswered. For example, if a confession excluded under section 76(2) is now considered to be worthless for *all* purposes, does it follow that such a confession is inadmissible for the purposes of challenging credibility *as well* as for the truth of its facts?

R v *Myers* in practice

As a final matter, it is instructive to consider the procedural operation of this new exception. Two particular aspects will be considered: defence applications to adduce a co-defendant's confession and how the jury should be directed upon the confession in the judge's summing up. In the context of defence applications to adduce a confession, it may be expected that the application will be made in one of three scenarios: first, the judge has ruled that the confession is admissible for the prosecution; secondly, the judge has excluded the confession for the prosecution; and finally, the judge has not as yet made any ruling, the prosecution having elected for whatever reason not to use the confession.

If the judge rules that the confession is admissible for the prosecution then the defence can obviously use the confession as part of its case. If the judge excludes the confession for the prosecution then the application of *R* v *Myers* in this context would depend upon the precise reason *why* the confession was excluded. If it was excluded as having been obtained under any of the circumstances specified in section 76(2) then the defence would be unable to use the confession: it would be 'worthless for all purposes'. If however the confession was excluded under section 78 then different considerations would apply. The confession, so long as it was relevant to the defence, would be admissible. Lord Hope has left the door ajar for trial judges to exclude the confession, but the precise grounds upon which this could be done are unclear. Finally, if the trial judge has not been required to make any ruling, the prosecution having elected for whatever reason not to rely upon the confession, then a determination would need to be made of whether the confession was obtained in a manner contrary to section 76(2); if so, the defence would be unable to use the confession.[37]

Turning to the judge's summing up, where a defendant is allowed to use a confession which is not relied upon by the prosecution or which is excluded under

35 [1997] 3 WLR 552 at 572.

36 *Lobban* v *The Queen* [1995] 1 WLR 877.

37 It is to be noted that in its report on criminal hearsay, published a matter of weeks before the decision of the House of Lords in *R* v *Myers*, the Law Commission recommended the inclusion in the Police and Criminal Evidence Act 1984 of section 76A, which would be analogous to section 76 and permit a defendant to adduce in evidence a co-defendant's confession. Section 76A(2) would provide that, if it is represented that the confession was or may have been obtained oppressively or in circumstances conducive to unreliability, it is inadmissible at the defendant's behest unless he or she proves on a balance of probabilities that it was not so obtained. See Law Commission (Law Com No 245), *Evidence in Criminal Proceedings: Hearsay and Related Topics* (London: HMSO, 1997).

© The Modern Law Review Limited 1999

section 78, the jury will have to be directed that the confession is only evidence for that defendant and that they are not to take the confession into account when determining the guilt of the co-defendant. Lord Slynn acknowledged the risk that the co-defendant may be prejudiced by the jury's misuse of the evidence, but seemed to dismiss it.[38] It would have been preferable for Lord Slynn to have acknowledged more openly the existence of a tension between the defendant's interest in presenting exculpatory evidence and the co-defendant's interest in not being unfairly prejudiced, and to have emphasised the consequent importance of giving the jury a carefully worded direction.

Conclusion

Prior to *R* v *Myers*, the House of Lords had been presented with the opportunity to reform the hearsay rule in *Myers* v *DPP*[39] and *R* v *Kearley*.[40] On both occasions, judicial hands were thrown in the air in favour of leaving any reform to Parliament. In contrast, *R* v *Myers* represents an approach of effecting reform by stealth. As this note has demonstrated, however, resort to the use of a 'hearsay-fiddle' carries with it the risk that new concepts are only half explained, and that the impact of new developments on existing law is not fully considered. This leaves uncertainty and unanswered questions. All of this could have been avoided if their Lordships had given bolder, clearly reasoned speeches making it plain that orthodox evidence doctrine was being turned on its head in a number of crucial respects. Only when we have such clear statements from the House of Lords can it finally be appreciated that the judiciary does have a vital role to play in reshaping and reforming the law of evidence, and that the notion that fundamental reform can be effected only by Parliament is simply wrong.

38 [1997] 3 WLR 552, 564.
39 [1965] AC 1001.
40 [1992] 2 AC 228.

© The Modern Law Review Limited 1999

REVIEW ARTICLE

Secrets of the Political Constitution

*Ian Leigh**

Adam Tomkins, **The Constitution After Scott: Government Unwrapped,** Oxford: Clarendon Press, 1998, xiii + 275 pp, hb £45.00, pb £18.99.

Sometimes the past seems not so much another country as another galaxy. After years of deprivation of real developments in the UK constitutionalists now have a surfeit: like the proverbial No 27 bus, devolution, the Human Rights Act, reform of the House of Lords, electoral reform and freedom of information have all come along at once. It is easy to forget that just three years ago the major source of interest was a drawn-out and convoluted investigation into an obscure area of government policy: export licence applications. The Major administration was overshadowed by *Waiting For Scott.* Unlike Godot, the central character finally arrived (in February 1996)[1] and for a fortnight the broadsheets and the sun-dried tomato buying classes were occupied with little else. For most normal life then resumed, but a faithful band of retired civil servants, political scientists and public lawyers pursued the lasting significance of it all. A veritable academic cottage industry – conferences, symposia, and special journal issues[2] – flourished. Adam Tomkins' book, building on his own earlier work,[3] is the latest contribution to this feast.

After a section setting the scene, the book is divided into four parts. The first three deal, respectively, with the position of government and Parliament, government and the intelligence agencies, and government and the courts. Part Four is a comparison of the Scott Inquiry with the Congressional investigations into Iraqgate. This framework allows for separate chapters on the subjects of ministerial responsibility to Parliament, the role of the civil service, freedom of information, the intelligence input to the export licensing process, and public interest immunity. The Conclusion assesses the Inquiry's significance in the context of constitutional reform.

Government, parliament and information

Although the Scott Report found that there had been repeated failures by Ministers to give an account to Parliament, paradoxically no one resigned in the aftermath.

* Department of Law, University of Durham.

1 *Report of the Inquiry into the Export of Defence Equipment and Dual-Use Goods to Iraq and Related Prosecutions*, HC (1995–96) 115, discussed in I. Leigh and L. Lustgarten, 'Five Volumes in Search of Accountability: the Scott Report' (1996) 59 MLR 695.

2 *Public Law*, Autumn 1996, *Parliamentary Affairs*, January 1997 (re-issued as F. Ridley and B. Thompson (eds) *Under the Scott-Light: British Government Seen Through the Scott Report* (Oxford: OUP, 1997)).

3 'Public Interest Immunity After Matrix Churchill' [1993] PL 650; 'A Right to Mislead Parliament?' (1996) 16 LS 63; 'Government Information and the Public: Misleading by Design or by Default?' [1996] PL 472; 'Intelligence and Government' (1997) 50 Parl Affs 109

© The Modern Law Review Limited 1999 (MLR 62.2, March) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

The chapter on Ministers and Parliament (ch 1) places Scott's strictures on ministerial responsibility in the continuing debate between the government and select committees over the distinction (if any) between 'accountability' and 'responsibility'. The history of the expression 'knowingly mislead' which occurs in the *Ministerial Code* (then called *Questions of Procedure for Ministers)* and by which Scott set such store is carefully rehearsed and criticised. The reader is also treated in the process to a valuable dissection of the policy-operations distinction and the controversy surrounding the sacking of the Director of the Prison Service, Derek Lewis, by Michael Howard in 1995. There is a worthwhile excursion to an earlier, and often misunderstood, constitutional *cause célèbre*, Crichel Down. This yields the lesson that Sir Thomas Dugdale's resignation has been incorrectly attributed to acceptance of responsibility for the actions of civil servants whereas it was to appease his own back-benchers in the Conservative Party who disliked the policy of keeping farms under state control. Equally misunderstood was Sir David Maxwell-Fyfe's famous typology of four different situations in which ministers are responsible to Parliament as regards the actions of civil servants.[4] Tomkins persuasively argues that the purpose was to *deny* the existence of any actions of their departments for which ministers would not be 'responsible', in the face of contemporary fears about anonymous bureaucracy. It is therefore a defective foundation for the distinction between 'responsibility' and 'accountability' which the Conservative governments and senior civil servants later tried to build upon it.

Scott, however, accepted this distinction uncritically, due, Tomkins argues, to constitutional naivety and his failure to appoint expert assessors familiar with the substantial debate over accountability which had been raging since Westland (p 64). Nevertheless, the report played an important part in the debate which led ultimately to Parliament adopting resolutions in the dying hours before prorogation for the 1997 general election which set out its own understanding of the doctrine.[5] These resolutions record the duty to account (including for the actions of Next Steps agencies), stress the need for 'accurate and truthful information', that information should only be withheld when in the public interest, and that ministers should require civil servants to be as helpful as possible in providing accurate, truthful and full information to select committees. They assert that inadvertent errors should be corrected at the earliest opportunity and that ministers who knowingly mislead Parliament will be expected to offer their resignation to the Prime Minister. Although the content of the resolutions was a political compromise, nevertheless, Tomkins suggests, this is a significant breakthrough:

> No longer is ministerial responsibility merely an unwritten constitutional Convention. ... It is now a clear parliamentary rule, set down in resolutions by both Houses of Parliament. ... the government acting on its own cannot now change the terms on ministers' responsibility to Parliament in the way that the Conservative government did throughout its period in office (p 62).

Important as this is, it is only part (and perhaps a small part) of what is needed to make ministerial responsibility *work*. The more difficult question is political will. The conclusion that the doctrine of ministerial responsibility 'is in a healthier condition now than it was at the beginning of the 1990s' (p 64) depends crucially on a calculation of probabilities. Tomkins is optimistic about the future and suggests that 'the Scott report demonstrates the failure, but also the hope, of

4 HC Deb vol 530 cols 1285–1287 20 July 1954.
5 HC Deb vol 292 cols 1046–1047 19 March 1997; HL Deb vol 579, col 1055 20 March 1997.

© The Modern Law Review Limited 1999

Parliament' (p 66). Others may be less sanguine about Parliament's capacity to engage in disinterested enforcement of the convention. Perhaps as a result of his optimistic conclusion, the chapter devotes little space to exploring mechanisms by which, in his words, 'ministerial responsibility can return to its proper status as an effective weapon in parliament's armoury by which ministers are held fully to account and are made truly, and constitutionally responsible' (p 67).

Overall, then, the account is too rosy: in recent constitutional crises there have been four basic difficulties over ministerial responsibility – agreeing the content of the convention, investigating the facts, agreeing on the application of the convention to the facts, and enforcement. We are now nearer (in an admittedly somewhat unsatisfactory way) to resolving the first of these difficulties, but the other three remain intractable.

The chapters on the civil service and on freedom of information face the difficulty that the Scott report, to the disappointment of a number of commentators, made no formal recommendations about these matters. Nevertheless, as Tomkins correctly identifies, concerns of this kind were central to the investigation, albeit not treated explicitly. In each case, as with ministerial responsibility, he argues that it should be for Parliament and not the government to set the ground rules. He is, therefore, in favour of a statutory regime for the civil service[6] incorporating protections for whistle-blowers. 'The cloak of the Crown, so convenient to government, should be stripped away, and the civil service should be re-clothed in a new parliamentary garb' (p 94). Likewise, 'the public interest is not a matter only for ministers, but should be enshrined in statute to include what might be called basic constitutional values' (p 94). In the same way, after criticising the revised rules governing the evidence of civil servants to parliamentary select committees (*Departmental Evidence and Response to Select Committees*, 1997), he urges Parliament to devise rules of its own (p 111).

Statutory freedom of information is to be preferred to the Code of Practice on Access to Government Information for the same reason (p 125), although his account also acknowledges the credit due to the Major administration for introducing the Code, the generally beneficial way in which it has operated, and the important part played by the Ombudsman. The discussion is measured and thoughtful, although it has inevitably been overtaken by publication of the Government's White Paper, *Your Right to Know*, detailing plans for a Freedom of Information Act.[7] This makes it possible to predict with greater certainty an issue which Tomkins raises but, disappointingly, never fully tackles: whether freedom of information legislation would have averted the Arms to Iraq affair in the first place.[8]

Although it proposes a reduction from the 15 exemptions in the Code of Practice, the White Paper contains no shortage of proposals for exemptions which could have been used to withhold information. The proposed seven specified interests include: national security, defence and international relations; information the disclosure of which could harm the effectiveness of law enforcement; commercial confidentiality; the safety of the individual; and information supplied in confidence

6 cf N.D. Lewis, 'A Civil Service Act for the United Kingdom' [1998] PL 463.

7 Cm 3818 (1997). For discussion see: Birkinshaw, [1998] PL 176; S. Palmer 'Freedom of Information – Principles and Problems: A Comparative Analysis of the Australian Proposed UK Systems' in University of Cambridge Centre for Public Law, *Constitutional Reform in the United Kingdom: Practice and Principles* (Oxford: Hart Publishing, 1998).

8 see 115 and 124. There is, however, a useful treatment of US Classification and FOIA exemptions (248–254) which he uses to reach a broadly negative conclusion (at 264).

© The Modern Law Review Limited 1999

by private individuals and companies or other organisations.[9] In all these cases the harm to the specified interest would need to be a 'substantial' before it would qualify for protection, and it would also be necessary to establish that it was in the public interest to withhold the information. (However, the White Paper endorses in passing the so-called 'mosaic' argument frequently invoked by intelligence agencies when it states 'such harm could in certain circumstances be caused by the cumulative effect of numerous disclosures of similar material over a period, as well as by a single disclosure').[10] A large proportion of the information disclosed by the government to the Scott Inquiry, but previously withheld from Parliament and the public, about export licence applications, defence procurement and Iraqi attempts at proliferation would fall into these categories. The discussions of officials and ministers, whether based on information from these or other sources, concerning export licence applications would certainly have fallen under a proposed exemption for policy advice for which the test would be 'simple' rather than a 'substantial' harm.[11] All the familiar arguments about the 'maintenance of collective responsibility in government', 'the political impartiality of public officials' and 'the importance of internal discussion and advice being able to take place on a free and frank basis' are summoned in support of this. It is as though the Scott Inquiry had never taken place.

If all of this were not enough, the White Paper proposes that two relevant areas should be not merely exempted but *excluded* altogether from the legislation (with the consequence that the Information Commissioner would have no jurisdiction). The security and intelligence agencies, together with information about these organisations held by other public bodies, are the first. The assertion that these bodies 'could not carry out their duties effectively in the interests of the nation if their operations and activities were subject to freedom of information legislation'[12] has rightly been questioned by the Public Administration Committee in its review of the Government's proposals.[13] As the Committee pointed out, even if the agencies' work were to be covered by the legislation, in practice they would frequently succeed in invoking exemptions to defeat access requests. Nevertheless, there is an important symbolic point at stake: requiring a public body to justify withholding information is very different to exempting it altogether. It is noticeable that the Data Protection Act 1998, dealing with access to personal files, follows the route of exemption and not exclusion. A minister may issue a certificate under section 28 of that Act claiming national security as a reason for exemption, but the certificate is reviewable by the Tribunal. Empirical research on the operation of Canadian Access to Information Act 1982 and the Privacy Act 1982 concerning access to intelligence files confirms the benefits of subjecting the information-handling processes of such agencies to outside scrutiny by an independent official, not least in stimulating internal awareness of information and privacy concerns.[14] The second exclusion relates to law enforcement and is intended to cover not only the police but also other bodies carrying out investigations, prosecution or

9 *Your Right to Know*, para 3.11.
10 *ibid* para 3.10.
11 *ibid* para 3.12.
12 *ibid* para 2.3.
13 *Third Report of the Public Administration Committee, Your Right to Know: the Government's Proposals for a Freedom of Information Act*, HC 398-I (1997–89) paras 38 and 39.
14 Interviews were conducted with officials of the Canadian Security Intelligence Service, the Information and Privacy Commissioners and their staff, users of the legislation, Federal Court judges and other experts. see L Leigh, 'Legal Access to Security Files: the Canadian Experience' (1997) 12 *Intelligence and National Security* 126.

© The Modern Law Review Limited 1999

regulatory work. Perhaps wisely, the White Paper did not attempt to list these, but it would be surprising if Customs and Excise (prosecutor of the Matrix Churchill directors) were not included. Again the Select Committee concluded that a wholesale exemption of this kind was unnecessary and recommended that law enforcement agencies be able to take the benefit of an exemption based on a simple harm test.[15]

Turning from the Labour government's words to its actions, the picture is equally mixed. The admission that the White Fish Authority (a quango much-parodied by an earlier generation of administrative lawyers) was after all a cover for a naval intelligence operation based in Grimsby suggests a hitherto disguised talent for irony in Whitehall. The opening of the Security Service Website can be seen as post-ironic.[16] The government's strongest claim to the freedom of information mantle was the Home Secretary's announcement of the numbers of files held by the Security Service.[17] However, the same statement rejected any claims that individuals may have to access to their own files, however old, confirmed that the Service had embarked upon a programme of destruction of these files without independent scrutiny, and rejected any form of outside control. Equally, the attempts to extradite from France the former Security Service officer David Shayler and the government's refusal to engage in substantive discussion of his allegations of incompetence and impropriety by the intelligence agencies even after his successful invocation of the 'political offence' exception[18] is reminiscent of the crusade against Peter Wright.

A major test for the Blair administration's commitment to freedom of information in a post-Scott environment, its relations with Parliament and of its new 'ethical' foreign policy[19] came in a scandal bearing an uncanny resemblance to the Arms to Iraq affair. The question was the extent of British involvement in the restoration to power of President Kabbah in Sierra Leone following the overthrow of the military junta in March 1998 – either directly or through connivance at the actions of a British-based company specialising in supplying 'military services', Sandline International. As in the Scott inquiry, allegations were made that ministers had misled Parliament over the state of government knowledge in

15 Public Administration Committee, paras 24–30

16 http://www.mi5.gov.uk

17 HC Deb vol 317 cols 251–254 29 July 1998 Of 725,000 files opened since 1909 by the Service 285,000 have been destroyed (110,000 of these in early 1990s). The current holding of 440,000 files includes 290,000 personal files of which 20,000 are active; 13,000 of these relate to UK citizens (divided approximately equally between terrorism and the Service's other current interests, espionage weapons-proliferation and serious crime). The greatest controversy concerns the holdings of inactive counter-subversion files compiled during the Cold War. Further details of the criteria for the destruction of files were given in an earlier statement HC Deb vol 305 col 520 20 January 1998, see also *MI5: The Security Service* (London: HMSO, 3rd ed, 1998) 24–25 The Lord Chancellor's Advisory Committee on Public Records has been asked to advise on the criteria for retention of files on grounds of historical interest: *Government Response to the Annual Report of the Intelligence and Security Committee for 1997–98*, Cm 4089 (1998) para 16. During the debate on this report the Home Secretary undertook to look again at the whole question· HC Deb vol 318 col 649–650 3 November 1998. The Intelligence and Security Committee has also undertaken to examine whether an individual should have rights in connection with the destruction of files, the protections against inaccurate information, the application of data protection legislation, and implications of the ECHR· *Annual Report of the Intelligence and Security Committee for 1997–98*, Cm 4073 (1998) para 51. The same report discloses that the Secret Intelligence Service (MI6) holds 86,000 files (75 per cent are closed) of which approximately half relate to UK citizens, although they are not files whose subject-matter is individuals· *ibid*, para 52.

18 For an account by his supporters see: http://www.shayler com

19 Designed in part to make human rights considerations an explicit factor in decisions such as the grant of export licence applications: L Lustgarten, 'The Arms Trade and the Constitution· Beyond the Scott Report' (1998) 61 MLR 499, 510, citing HC Deb vol 299 cols 26–29 (WA) 28 July 1997.

© The Modern Law Review Limited 1999

answers given by the junior Foreign Office ministers Baroness Symons[20] in the House of Lords and Mr Lloyd in the House of Commons.[21] By a bittersweet irony two of the main protagonists, the Foreign Secretary, Robin Cook, and Baroness Symons, had been involved in the Arms to Iraq affair, the former as tormentor in chief of the government from his position as Shadow Trade Secretary, and the latter as defender of senior civil servants whilst Secretary to the First Division Association. Another common theme was allegations of connivance at illegality: in this case an apparent breach of the Sierra Leone (United Nations Sanctions) Order 1997, which prevented the supply of military equipment to either side. Sandline's solicitors objected to a Customs investigation into the actions of their clients, arguing that any breach had been 'licensed' through contact with the High Commissioner, Mr Penfold.

An inquiry conducted by Sir Thomas Legg and Sir Robin Ibbs exonerated ministers from any intention to mislead and blamed inaccuracies on poor briefing by civil servants and failures of communication.[22] Whilst not finding, as had been alleged, that the High Commissioner had in effect arranged the military intervention, it nevertheless recommended clearer guidelines for contact between Foreign Office staff and military companies and that greater attention be paid to the requirements of domestic law (Foreign Office officials were said to be too preoccupied with international law). The treatment of this report suggested that the new government intended to continue the practices of the previous administration in attempting to manage news and Parliament. Speculation about the Legg Report had dominated the news for weeks beforehand but it received comparatively little coverage when published since the Foreign Secretary's statement to the House of Commons fortuitously coincided with the first Cabinet reshuffle of the new government – in which the major casualty was the minister with responsibility for freedom of information legislation, Dr David Clarke!

Attempts by the Foreign Affairs Select Committee to investigate the affair have brought the first real clashes between the new government and Parliament over select committees' access to information. The Committee vigorously contested the Government's view that it should await the outcome of the Legg report before proceeding with its own investigation. The Foreign Secretary's denial of access to Foreign Office telegrams led the Committee to issue two special reports, with the result that they were given a summary of the telegrams and the opportunity for some members to check the summary against the originals.[23] The appearances of the Permanent Secretary before the committee resulted in Sir John Kerr being forced to correct earlier evidence which he had given about briefings of ministers. His refusals to answer questions led to threats of contempt proceedings. Controversially, Labour members of the committee blocked some questions from other members by formally objecting to them. The select committee's investigation is continuing at the time of writing. Nevertheless, on a provisional assessment the episode seems to demonstrate that the new government is as keen as its predecessors to control the flow of information to Parliament, but also that, post-

20 HL Deb 587 cols 99–101 10 March 1998.

21 HC Deb vol 308 cols 840–846 12 March 1998.

22 *Return to an Address of the Honourable the House of Commons dated 27th July 1998 for the Report of the Sierra Leone Arms Investigation*, HC 1016 (1997–98); see esp. ch 9; for the Foreign Secretary's statement and subsequent questions see HC Deb vol 317 cols 19 ff 27 July 1998

23 *First Special Report of the Foreign Affairs Select Committee*, (1997–98) HC 760 and *Second Special Report of the Foreign Affairs Select Committee*, (1997–98) HC 852. It is unclear from the exchanges so far published whether the Foreign Secretary has acceded to the Committee's reported request to interview the Chief of the Secret Intelligence Service (MI6).

© The Modern Law Review Limited 1999

Scott, parliamentary committees are becoming less quiescent. In their case at least Tomkins' hopes for the 1997 resolutions may be partially vindicated.

Intelligence and accountability

Whereas historians have long recognised the importance of intelligence in understanding international and domestic political events, other disciplines have been slower to appreciate the central and routine part that it plays in government. Although the Scott report does not deal with this as an explicit theme, it is, nevertheless, the most detailed case study available on the place of intelligence within Whitehall. The report refers copiously to contacts between SIS and the Security Service and with their sources in Matrix Churchill, Paul Henderson and Mark Gutteridge, and quotes extensively from Contact Reports compiled following each interview with them and from intelligence reports circulated by the agencies. No official document has done so in Britain before and it is noticeable that, despite having access to similar material, the Legg report on Sierra Leone treats it with the conventional secrecy.

Chapter 4 ('Governing Without Intelligence') is a review of this neglected aspect of the Scott report – the light it sheds on the performance of the intelligence agencies and on the assessment and distribution of secret intelligence as an aid to government decision-making. Ex-ministers are tellingly quoted about the quality of intelligence (pp 154–155). David Mellor revealingly confesses that intelligence reports 'didn't tell you all you wanted to know about life'. Lord Howe's (somewhat crisper) assessment is that they were 'cornflakes in the wind'. Tomkins' claim (p 162) that the Report had less impact on secret intelligence than any other area of government activity is not readily verifiable, despite the various announcements of reforms which have been made.

He also criticises the Intelligence and Security Committee (established under the Intelligence Services Act 1994)[24] for doing too little to follow up leads from the Scott report (pp 159–162). However, it was to be expected that the Committee would take some time to establish its authority in the intelligence community and it made its appearance at a time of extraordinary realignment of security and intelligence priorities, which naturally had a strong prior claim on its time. The Committee's progress must be measured against the relative lack of legal powers which it enjoys and lack of an investigative capacity. Judged by these standards the Committee appears to be performing strongly. It has flexed its muscles by publicly asserting that it expects to be 'properly and promptly informed' by the agencies of their activities, rather than merely responding to requests for information; in this the Committee were consciously following the Congressional oversight model, rather than the more responsive mode contemplated in the legislation.[25] The agencies appear to be co-operating in this process. Where the Committee can perhaps be criticised is for a cautious approach to publishing details of its investigations. There are signs, however, in the most recent report (the first to be debated in the House of Commons)[26] that the Committee, in part due to the Shayler affair, is becoming more vocal in its criticisms and is attempting to enlist public and parliamentary support in a quest for greater powers.

24 L. Lustgarten and I. Leigh, *In From the Cold: National Security and Parliamentary Democracy*, (Oxford: Clarendon Press, 1994), 511–515.

25 *Annual Report for 1995*, Cm 3198 (1996) para 37.

26 HC Debs vol 318 cols 578 ff 3 Nov 1998.

© The Modern Law Review Limited 1999

The Committee can be criticised also for failure to pursue some problems relevant to its previously published concerns over the use of intelligence agencies for law enforcement.[27] Despite clearly cataloguing conflicts of interest in the intelligence agencies' handling of human sources (Henderson and the Ordtec Managing Director, Paul Grecian) between obtaining intelligence and passing on information for enforcement of export control legislation, Scott made no recommendations. These difficulties are likely to multipy as the Security Service takes a more prosecution-oriented role. As *Justice* has recently observed, the whole area is ripe for greater regulation.[28]

The new remit given in the Security Service Act 1996 of assisting in the prevention and detection of serious crime, together with the increasing tendency for intelligence officials to appear in court as part of MI5's lead role against terrorism means that the vexed topic of judicial treatment of secret information has also assumed fresh significance. Many lawyers will be familiar already with the account of public interest immunity given in chapter 5: it traces the development of the doctrine through the landmark cases,[29] the use of PII certificates in the Matrix Churchill trial, and the Report's criticisms of the extended use of class certificates, the application of PII to criminal trials, and of the Government's claim that ministers *had* to sign to make these claims. The development of PII in *Wiley*[30] is reviewed, as is (briefly) the change in practice announced by the government in December 1996.[31] Overall, Tomkins consistently prefers Scott's interpretation of the law to the government's and roundly criticises the Attorney-General's ill-judged initial response to the Report. Competent as a review of the law, the account in this chapter is, nevertheless, like the Report itself on this issue, disappointing. Like Scott before him, Tomkins rather loses sight of the bigger constitutional picture underlying the technicalities of PII. The important questions about the separation of powers and the rule of law are glimpsed now and then through the thicket of the detail, but not directly confronted.[32] In view of the book's emphasis on the lasting significance of the Report, a more critical analysis might also have been expected of the Government's change of practice announced in December 1996, and more attention paid to the new regime for disclosure.

The latter point is particularly important for two reasons. First, the Criminal Procedure and Investigations Act 1996 now provides the procedural context in which all PII claims are made in criminal cases and, routinely, in relation to informers. Secondly, the more restricted statutory duties of prosecution disclosure under the Act mean that defendants are much less likely now than at the time of the Matrix Churchill trial to be able to claim either primary or secondary disclosure of unused prosecution material. In effect, the Crown has much less reason to claim PII since it is less likely to be under a duty to disclose in the first place. (Discussion of this important development is relegated to a footnote on p 190).

It is to be expected also that in future the development of PII will be influenced

27 *Report on the Security Service's Work Against Organised Crime*, Cm 3065 (December 1995).

28 Justice, *Under Surveillance: Covert Policing and Human Rights Standards* (London: Justice, 1998), esp ch 2.

29 *Duncan* v *Cammel Laird* [1942] AC 624; *Conway* v *Rimmer* [1968] AC 910; *Burmah Oil* v *Bank of England* [1980] AC 1090; *Air Canada* v *SS for Trade (No 2)* [1983] 2 AC 394.

30 *R* v *CC of West Midlands, ex p Wiley* [1995] 1 AC 274.

31 HC Deb vol 576 cols 1505–1517 19 December 1996. For comment see M. Supperstone, 'A New Approach to Public Interest Immunity?' [1997] PL 211; C. Forsyth, 'Public Interest Immunity: Recent and Future Developments' (1997) 56 CLJ 51.

32 For an account taking this perspective, see I. Leigh, 'Public Interest Immunity' (1997) 51 *Parliamentary Affairs* 55.

© The Modern Law Review Limited 1999

by Article 6 of the European Convention on Human Rights.[33] At Strasbourg, recent decisions of the European Court of Human Rights suggest an increasingly interventionist approach towards the right to a fair trial.[34] In *Tinnelly and McElduff* v *UK*[35] a provision of the Fair Employment Act 1976 in Northern Ireland enabling a minister to give a conclusive certificate that the reasons for a decision were national security was held to violate Article 6 on the grounds that it prevented the tribunal from effectively reviewing the facts. Although PII certificates are no longer conclusive in the same way as the contested certificate, it is not hard to see how, when combined with the European Court's new preparedness to review the overall conduct of a trial for fairness, their use might be open to Convention challenge. One case pending in Strasbourg (a complaint brought following *R* v *Davis, Johnson and Rowe*)[36] may yet call into question the whole new statutory regime for disclosure in criminal cases which, like that controversial ruling from the Court of Appeal, allows the prosecution in some instances to apply for directions without advising the defence.

Iraqgate and Congress

Comparative analysis is one of the book's original contributions to discussion of Scott. *The American Connection* (Part IV) will be unfamiliar to most British readers. The examination of how successfully US constitutional mechanisms grappled with very similar issues to those raised before the Scott inquiry provides a unique opportunity for comparative constitutional study. In a readable and a valuable account of a complex story, Tomkins unpicks the events surrounding the investigation of the Gonzalez Committee into the financing of Iraqi weapons procurement, attempts by the Bush administration to frustrate these, and the prosecution of Christopher Drogoul, the manager of the Banca Nationale del Lavoro (BNL) Atlanta branch. As Tomkins notes, effective investigation was hampered by the fact that no single independent investigator had responsibility to consider all aspects of the affair. The Gonzalez Committee was the House Banking Committee, and so had a constitutionally limited role: its central concern was the apparent failure of the federal banking regulation system to monitor or control lending and credit guarantees given to Iraq. Apart from BNL, the other major institutions of interest to Gonzalez were Eximbank, a governmental agency with the statutory object of aiding and facilitating trade through credit guarantees, and what was ostensibly a government agricultural credit guarantee programme, the Commodity Credit Corporation. Each of these had massive exposure to Iraq. The Gonzalez Committee faced numerous attempts to obstruct its investigation by withholding information, including claims that the release of information to it would prejudice the forthcoming Drogoul prosecution or might compromise national security. Some documents disclosed were heavily redacted and there was

33 For an initial tentative assessment see Leigh, *ibid* 62–64.

34 See especially in a criminal context *Barbera, Messegue and Jabardo* v *Spain* (1988) 10 EHRR 87, holding that the Court can review the whole treatment of the evidence by the domestic judge. In *Van Mechelen* v *Netherlands* (1998) 25 EHRR 647 the court found a breach of Art 6 because there was no evidence that the domestic court had correctly weighed the prejudice to the defence in allowing cross-examination of anonymous witnesses via a video link which did not allow observation of demeanour. See also *Teixera de Castro* v *Portugal*, ECHR, Judgment 9 June 1998 (conviction on basis of evidence from police officers who were agents provocateurs violated Art 6).

35 *Tinnelly and McElduff* v *UK*, ECHR, Judgment, 10 July 1998; for earlier discussion see K. McEvoy and C. White, 'Security Vetting in Northern Ireland. Loyalty, Redress and Citizenship' (1998) 61 MLR 341, 349–354.

36 (1993) 96 Cr App R 110; *Rowe and Davis* v *UK* (1998) 25 EHRR CD 118, admissibility decision.

© The Modern Law Review Limited 1999

a systematic attempt from the White House to co-ordinate and delay the release of information to the Committee, orchestrated by a group of legal advisers who became known as the 'Rostow gang'.

British constitutional commentators often assume that Congressional committees are in a superior position to Select Committees because of the ability to issue subpoenas and high levels of support staffing. However, as Tomkins explains in his review of the constitutional precedents and case law, a number limitations hamper their work: the requirement that Congress establishes a legitimate *legislative* interest in the matters under investigation (rather than exposure of Executive wrong-doing as such), the constitutional rights available to private individuals called before congressional investigations, the contentious use of executive privilege, and the classification of documents under the Freedom of Information Act. A strong and interesting theme running through this account of the US constitutional practice is the extent to which conflicts between the Congressional and the Executive branches over access to information have repeatedly been resolved by negotiation rather than by determination in the courts.

In passing, chapter 7 (at pp 262–263) discusses the debates concerning the role of Special Prosecutors (Independent Counsel appointed by the Attorney-General under the Ethics in Government Act). None was appointed to investigate Iraqgate, but with hindsight and in view of the controversy surrounding the Starr report into President Clinton,[37] it would have been interesting to have seen a fuller appraisal of this constitutional device by way of comparison with the Scott report or the Parliamentary Commissioner for Standards (perhaps the nearest that the British constitution comes to such an office). Sir Richard Scott,[38] Kenneth Starr and Sir Gordon Downey could each argue with some justification that attacking the investigator is a well-tried diversionary tactic employed by politicians under pressure.

Ultimately, Congressional interest in the investigations petered out. After Clinton became President in 1992 there was little further political point in pursuing Iraqgate and the Democrats lost control of the House of Representatives in the 1994 elections in any event. Nor could either of the Congressional Intelligence Committees be persuaded to open formal investigations into the scandal, and they alone had power to obtain access to the necessary secret intelligence and national security documents. Surprisingly, therefore, more of the story was uncovered in Britain than in the US. At the very least, Tomkins argues, this must give advocates of a written constitution and supporters of constitutional reform, such as freedom of information, pause for thought.

Conclusion: reforming the political constitution

The conclusion reflects a recurring theme of the book: Tomkins' indefatigable belief in political accountability as preferable to legal accountability. In part this seems based on incurable optimism, like the drinker whose glass is half-full, rather than half empty. And so he finds hope in the behaviour of the two Conservative MPs (Quentin Davies and Richard Shepherd) who voted independently in the debate on the Report on 26 February 1996, rather than the 320 who did not (pp 269–271). This points the way to the 'constitutional ideal which it is not only

37 *Referral to the United States House of Representatives pursuant to Title 28, United States Code, § 595 (c) Submitted by the Office of the Independent Counsel September 9, 1998*

38 Controversy over the inquiry process is discussed by Tomkins at 17–21.

© The Modern Law Review Limited 1999

desirable but essential that we move': breaking the domination of Parliament by the political parties. Proposals for providing a career structure for MPs by giving them senior responsibility for scrutiny of the executive and of legislative proposals seem worthwhile avenues as an alternative to the scramble for ministerial office, which leaves MPs overly-beholden to the Whips. Tomkins therefore endorses the proposals from the Liberal Democrats for revitalising select committees by establishing a separate parliamentary pay-roll for their members and those who would chair them, and through introducing a rapporteur system.[39] These are valuable suggestions, but it is worth also considering other mechanisms which might, to use his expression, 'drive a wedge between government and Parliament'. Some type of standing, independent, fact-finding office to build on the models of the Ombudsman and the Parliamentary Commissioner for Standards, or to check the veracity of answers to Parliamentary Questions,[40] could also help to provide MPs with the necessary data which is a precondition to scrutiny.

His preference for Parliamentary reform and reflection on US experience lead Tomkins to the following fork in the road:

> should we throw away the British political constitution, give up on Parliament, and turn instead to the courts; or should we seek to amend Parliament so that its place at the heart of the constitutional control room can be restored? (p 271)

This is, however, a false alternative. Certainly, the Bill of Rights debate from Griffith's celebrated Chorley lecture[41] onwards has led to a tendency to see judicial or political accountability as a zero sum game, but the broader range of constitutional reform proposals which are currently either being implemented or under discussion demonstrates the falsity of this. To take but one area close to Tomkins' subject-matter – freedom of information – a whole variety of enforcement mechanisms involving both political and judicial personnel can be envisaged.[42] At the judicial end of the spectrum there is the United States' model in which the courts have the power to review Executive refusals of access to information on appeal. In Canada an Information Commissioner investigates complaints against refusal of access but only the Federal Court has power to order disclosure; however, in practice the Commissioner plays a pivotal role in securing greater access by negotiation and it is extremely rare for a government department to ignore such recommendations. Under the Australian legislation it is possible in some circumstances for a minister to refuse to comply with a recommendation for disclosure, but the reasons for refusal must be delivered to the legislature, so inviting political scrutiny. The Labour government's White Paper suggested in effect building on the Ombudsman experience under the Code of Practice by creating an independent Information Commissioner to hear appeals and with power to order disclosure, subject only to judicial review.[43] Each of these different schemes has something to recommend it in its own constitutional setting and each

39 At 274, citing *A Parliament for the People: Proposals for Reform of the House of Commons* (policy paper No. 20 September 1996).

40 See Leigh and Lustgarten, n 1 above, 715 for a proposal along these lines.

41 J.A.G. Griffith, 'The Political Constitution' (1979) 42 MLR 1.

42 See generally R.S. Baxter, 'Freedom of Information: Dispute Resolution Procedures' (1996) 2 *European Public Law* 635.

43 *Your Right to Know*, Cm 3818 (1997), paras 5 16 and 5.12. However, the transfer of responsibility for the legislation to the Home Office has led to speculation that the proposed enforcement procedures may be diluted. Firm plans for the legislation for the Parliamentary session 1998–99 were markedly absent from the Queen's Speech. Subsequently the Home Secretary stated that he hoped the Bill would be published for consultation before the end of February 1999: HC Deb vol 321 col 569 30 November 1998.

© The Modern Law Review Limited 1999

involves subtle interplay of legal and political factors, which will in turn influence the political climate in which it operates. This only serves to illustrate the unreality of attempting to categorise such mechanisms as either political or legal.

Nor is the only choice between reforming Parliament and increasing the role of the courts. One of the intriguing features of the constitutional reform programme now underway is the extent to which power is being fragmented between different organs of the state. This is not merely due to the increased role for the courts under the Human Rights Act. It also involves the creation of subordinate regional legislatures in Scotland and Northern Ireland, executive devolution in Wales, the possible creation of regional assemblies for parts of England with a consultative role in relation to the new Regional Development Agencies, and the grant of experimental powers to local authorities and the emergence of new forms of local governance, such as directly elected mayors for major cities. In each case one could argue that the doctrine of Parliamentary supremacy has been preserved and central government will remain firmly in control of the purse strings (except in the case of the power of the Scottish Parliament to vary income tax rates). Nevertheless, for the first time in the post-war period we are now witnessing a voluntary reversal of the centrifugal tendencies of the state. If this results in a re-emergence of political pluralism and increased toleration of sources of non-central power, then it may in the end it may do more to check the powers of the over-mighty Executive than all the schemes one can dream up for the reform of Parliament. One should add to this picture the already complex process of fragmentation of power resulting from the creation of 'Next Steps' agencies, the growing band of regional and local quangos, the breakdown of formal boundaries between public and private agencies, and the direct and indirect influence of European institutions. Rhodes' thesis of the 'hollowing-out' of the state and the accompanying metaphor of policy networks[44] start to appear as more compelling ways of mapping the constitution than the old classifications of Parliament versus Government versus Judiciary, checks and balances, and so on. In retrospect, we may come to see the Scott Report and the debates it provoked as the last gasps of the old constitutional order.

44 R. Rhodes, 'The Hollowing Out of the State: The Changing Nature of the Public Service in Britain' (1994) 65 Pol Q 138.

© The Modern Law Review Limited 1999

REVIEWS

Hilaire McCoubrey, **The Obligation to Obey in Legal Theory**, Aldershot: Dartmouth, 1997, viii + 228 pp, hb £39.50.

This is an extremely frustrating book to read. The concluding paragraph provides a loose 'description of broad categories of elements associated with the phenomenon of "law"'; an observation that questions of obligation feature significantly within this loose description; and, in the final sentence of the book, the oddly tentative remark that 'the understanding of the association of obligation(s) with law is a core element of jurisprudential study and, thus, a case may be made out that the development of a modern theory of legal obligation may also be the key to a modern treatment of general legal theory' (p 218). This much could have been gleaned from an intelligent reading of the title accompanied by a fleeting acquaintance with the phenomenon of law.

It is true that, in the Foreword, the author reduces our expectations as to what the title may convey for this book. We are explicitly told not to anticipate 'some new substantive theory of "obligation"'; but the promise is nevertheless held out of a 'contextual analysis of the obligation(s) to obey', which will demonstrate the misconceptions underlying existing controversies, and deliver 'a means whereby a more fruitful analysis of a complex central legal phenomenon may be advanced' (p vii). This promise is repeatedly echoed in the comments found towards the end of each of chapters 1–7 (at pp 12f, 46, 86, 123, 157, 180, 199) which indicate the need to recognise the different categories, or concepts, or species, or forms of the obligation to obey law within a theoretical framework in which they can be related but also kept distinct.

These seven chapters commence with a briefer introductory chapter where the main credit for noting the central obligatory characteristic of law is given to H.L.A. Hart and the three categories of legal obligation are unwrapped – the coercive, the formal or intra-systemic, and the moral. There then follow four chapters surveying general theories of law in which the theorists' attitudes to legal obligation are sketched. This part of the book reads much like a jurisprudence textbook but the author's aim is to reveal the need to distinguish the different categories of obligation in order to obtain a clearer grasp of the more general jurisprudential differences exhibited. Next comes a case study of the obligation to comply with superior orders, which again stresses the importance of distinguishing different forms of obligation and recognising the possibility of conflict between them. The last of these seven chapters tackles the relationship between validity and obligation with particular reference to Nazi German law. Although the different forms of obligation have the potential to complicate the relationship between obligation and validity, McCoubrey settles on a concept of validity which is linked solely to the formal category of obligation and, as such, operates merely as 'a "bracketing" concept which identifies the cumulative effect of the application of formal criteria of identification' (p 199).

It is impossible to provide a fair assessment of these seven chapters without reading on to the final chapter, which is where we are presented with McCoubrey's 'Theory of Legal Obligation'; for it is in the promised exposition of the complex

© The Modern Law Review Limited 1999 (MLR 62.2, March). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

relationships between the different forms of obligation (eg, pp 12, 46, 180) that the full impact of these earlier studies can be appreciated. However, it should be said immediately that the first seven chapters of the book are unconvincing. The general jurisprudential material is eclectic, selected at times in a stimulating manner, but contains passages it would be disturbing to discover in a jurisprudence textbook – notably, the passing suggestion that Kelsen found 'little or no role' for obligation (p 5). Moreover, in its treatment of the jurisprudential material the book is inadequate. The debate between natural law and positivism is covered in complete disregard of the contributions made by Joseph Raz in *The Authority of Law* in 1979 and by John Finnis in *Natural Law and Natural Rights* in 1980, and in subsequent writings by these and other authors. The studies of superior orders and the relationship between validity and obligation are premised upon the particular approach to obligation(s) that the author takes without that approach being fully expounded. One must question the prudence of writing seven chapters of an eight chapter book which substantially depend upon a theoretical underpinning that is withheld until the final chapter. Keeping the reader in the dark until the last chapter may have worked for Agatha Christie, but it is less gripping in a work of legal theory.

The final chapter is a tremendous disappointment. The central analytical device employed by McCoubrey is a Venn diagram composed of three intersecting circles representing the three categories of obligation – coercive, formal and moral (p 211). This is accompanied by less than three pages of surrounding text which comment on the significance of the various sectors of the diagram, and little less than a page which seeks to illustrate how the insights conveyed by this analysis can be addressed to the different jurisprudential theories considered earlier in the book. The preceding and following parts of this chapter consist of highlighting points already covered in the previous chapters so as to reiterate the importance of establishing a theoretical framework in which the relationships between the three forms of obligation can be worked out, and the suggestion that a theory of legal obligation relates to the central issues of general jurisprudence. The actual exposition of the theoretical framework is accordingly limited to McCoubrey's Venn diagram and four pages of text.

The central sector of the diagram – where coercive, formal and moral overlap – is taken by McCoubrey to represent a Thomist conception of a complete or 'ideal' law, which 'satisfying formal criteria of identification is purposively appropriate and backed by the potential for coercive enforcement' (p 211). But McCoubrey does not in fact use the term law to signify the phenomenon captured in this sector of 'maximum interaction', preferring to call it a norm. This point of terminology raises a fundamental query concerning McCoubrey's diagram. It is without title – the three circles being independently termed 'moral obligation', 'coercive force', and 'formal obligation' – but he introduces it as depicting 'the three forms of obligation associated with law' (p 210). The inference from the introductory paragraph (and from the rest of the book up to this point) is that what we find within the various sectors are obligations relating to laws of various hues – in some way or another paler as we move out of the central sector which relates to the Thomist complete law. But the diagram itself has no label to signify this, and the commentary thereafter moves first from laws to norms, and then, in further sectors, on to the '"naked" phenomena' of moral obligation and coercive force which 'relate to a much wider field of action than that of positive law' (p 212). In other words, the diagram as a whole does not merely represent obligations associated with law.

© The Modern Law Review Limited 1999

In fact, the diagram does not merely represent obligations. One of the circles is entitled coercive *force* rather than obligation. The difference between obligation and brute force is acknowledged early on by McCoubrey (p 33) and the inappropriateness of the term obligation for coercive enforcement conceded in the final chapter (p 204). Yet this does not prevent McCoubrey from persisting with the trio of obligations throughout the book. Even in his exposition of the theoretical framework (pp 210–11), McCoubrey is content to vacillate between coercive force and coercive obligation, and so fails to confront what he appears to acknowledge elsewhere: one of the 'three forms of obligation associated with law' (p 210) is not a form of obligation at all.

These flaws in McCoubrey's analysis create an insuperable difficulty in relating sectors of the diagram to the legal phenomena they are supposed to represent. McCoubrey himself struggles to give an illustration for the sector comprising the moral-coercive overlap. Discounting a number of possibilities, he confidently asserts that '[e]xamples may readily be found in the past, for example under the mediaeval doctrine of the Two Powers' (p 212). But instead of providing an example, he leaves us to pursue the allusion by referring us back to chapter three. The closest scrutiny of pages 78–9, in which the doctrine is discussed, fails to yield any illustration of the phenomenon in question.

The search for an illustration of a law to which a ' "naked" formal obligation' applies (pp 212–13) involves a route that it is even more difficult to take than to describe. It commences not with an obligation but with a power, then moves on to an obligation resulting from the exercise of the power. It then requires the reader simultaneously to occupy the starting point where there is no obligation and hence no coercion, and the finishing point where the obligation is reached with eyes shut to the coercion that accompanies it.

The same difficulty faces McCoubrey in his attempt to relate the sectors of his diagram to the different perspectives within jurisprudential theory. He is not assisted by the introduction of '"fuzzy" borders marked by some degree of uncertainty', which purportedly accommodate the controversy over an obligation to obey a law 'on the edge of moral turpitude' (p 213). This confuses the difficulty of determining whether an obligation exists with the problem of what sort of obligation it is within McCoubrey's scheme. Nevertheless, on the back of this single confused instance, McCoubrey asserts that fuzziness exists between all the sectors.

Sadly, the fuzziness is not restricted to the regions between McCoubrey's sectors. The single extended paragraph on page 214 which peremptorily places all of the jurisprudential theories surveyed in earlier chapters into different parts of McCoubrey's diagram is so brazen in its construction as to require several pages of remedial work before it can be properly appreciated. Here, one example must suffice to illustrate the underlying flaws.

Hart's positivism is contrasted with 'classical' positivism (ie, Bentham and Austin) as falling into '"naked" formal obligation' as opposed to the sector of coercive/formal overlap. Putting aside any other problems concerning the relationship between coercion and obligation, the basic suggestion here that the development of positivist theory by Hart can be attributed to abandoning the coercive character of law is utterly unfounded. Yet McCoubrey is content to rely again on bald assertion where evidence is absent. In doing so, he misrepresents Hart's own endeavours to show how law is distinctive in being *more than* mere coercive force (see *The Concept of Law* (Oxford: Clarendon Press, 1st ed, 1961) at p 6).

McCoubrey's enterprise must be judged a failure in all material respects. It fails to represent obligations that are distinctively *legal* in scope; it fails even to

 © The Modern Law Review Limited 1999

elucidate *obligations*; and it confuses rather than illuminates different perspectives within legal theory. Is it possible for the consequent frustration to be tempered by some clues as to where more fruitful deliberation might lie?

Perhaps the most helpful departure from McCoubrey's approach would involve abandoning his search for three categories of obligation. We have already noted that McCoubrey himself is uneasy about treating coercion as a form of obligation, and in his adoption of the ideal Thomist law, coercion is the backing given to a law which 'carries' obligation (pp 210–11) rather than a form of obligation in itself. For McCoubrey, far from coercion being a category of obligation, quite the contrary holds: if there is an obligation to obey the law *then* coercion is justified to enforce it.

If McCoubrey's observations on the remaining two categories of obligation are considered, there arises the question of what exactly is the status of a formal obligation. In the context of a Thomist or a Natural Law perspective, the formal aspect serves a preliminary purpose in identifying an *obligation to obey*: it identifies the law, and then moral grounds provide an obligation to obey that law. In the straightforward case, the obligation to obey arises because the content of that law conforms to a moral standard. In the oblique case, where the content of the individual law fails to conform to a moral standard, the obligation to obey the law may still arise so as to ensure social stability – and this in itself is a moral ground. In neither case does the formal aspect itself identify an obligation to obey law: where neither of the two types of moral grounds is present, no obligation to obey exists – even though the formal aspect of law is present. There is, accordingly, no evidence yet of the existence of a formal obligation to obey law.

It is significant that when McCoubrey moves away from a Thomist or Natural Law perspective he recognises that a portrayal of a formal obligation to obey law is in danger of being circular, yet persists in the view that '[p]ositive law is ... a structure of norms which ... endows propositions with an obligatory character' (p 209). This is not so much circular as otiose – if positive laws *are* norms, then legal propositions do not need to be endowed with an obligatory character: they already possess it. The formal identification of a legal proposition in this positive sense is precisely an identification of a legal obligation – but not the identification of an obligation to obey the law. So neither here does McCoubrey present any evidence for the existence of a formal obligation to obey law.

If McCoubrey's search for three forms of obligation to obey law is futile, that is not to say that the underlying quest to confront the nature of legal obligation is a vain one. Nor is it to suggest that there is not any significance in considering the relationships that may exist between the formal identification of law, the potential for coercive enforcement of law, and any moral assumptions law rests upon. The suggestion is, rather, that whereas *legal obligation* may be associated with formal and coercive aspects of law, the question of whether there is an *obligation to obey law* cannot be analysed on the same level, since it opens up and subjects to evaluation the assumptions on which a formal or coercive analysis proceeds. The obligation to obey law opens up the assumptions on which law is based, but a theory of the obligation to obey law should also open up any further assumptions we make in evaluating those assumptions.

*Andrew Halpin**

* Faculty of Law, University of Southampton.

© The Modern Law Review Limited 1999

Andrew Halpin, **Rights and Law: Analysis and Theory**, Oxford: Hart Publishing, 1997, ix + 284pp, hb £35.00; pb £16.99.

Andrew Halpin's book is serious and intelligent, but extremely uneven; on the whole, its shortcomings somewhat outweigh its virtues. Amid the profusion of rebarbative postmodernist blather in contemporary legal theory, one would like to commend *Rights and Law* – a book of refreshingly straightforward and earnest philosophical analysis – without any major reservations. Alas, Halpin's arguments are too often disappointing to permit such a commendation.

As Halpin himself acknowledges, his ten chapters are rather loosely connected – a situation aggravated by the fact that his jejune introductory chapter bears quite a tenuous relation to the rest of the book. Chapters 2, 3, and 4 explore the key legal concepts of rights, powers, and claims; chapters 5 and 6 develop and apply a heuristic framework for discussions of rights; chapter 7 offers a critique of Roberto Unger's conception of rights; chapters 8 and 9 explore the debate between utilitarianism and theories of rights; and chapter 10 furnishes a concluding overview (which is more closely connected to the middle eight chapters than is the Introduction).

By far the weakest of the chapters is the second, on Wesley Hohfeld's analytical theory of rights. First published as a journal article in 1985, it is deeply and irredeemably confused. The book would have been immeasurably improved if this chapter (and any traces of it in the subsequent chapters) had been altogether expunged. No short review can straighten out all the muddles in Halpin's analysis of Hohfeld, but one central point – which Halpin invokes in some later chapters – will be discussed presently.

Chapters 3 and 4, on legal powers and claims respectively, are considerably better than the second chapter. Chapter 3, the best essay in the book, develops a conception of legal powers that is presented as a contrast to the conceptions propounded by Hohfeld and Joseph Raz. Halpin errs in thinking that he (or Raz) has pointed out mistakes in Hohfeld's approach to legal powers; his own stipulative conception of legal powers is merely an alternative to Hohfeld's stipulative conception, rather than a correction of errors therein. All the same, Halpin's discussion here is sophisticated and thoughtful. Chapter 4 is likewise interesting, or at any rate it becomes interesting after belabouring some fairly obvious points. Though it frequently uses ontological language to discuss epistemological issues, its basic thesis – positing an equivalence between rights and justified normative claims – is upheld convincingly against Alan White's contrary stance.

Chapter 5's heuristic framework is expounded rather tediously. (Throughout the book, Halpin's prose is plodding and sometimes inelegant, with numerous minor solecisms and fragmentary sentences.) Apart from some analytical lapses – such as a failure to distinguish clearly between conflicting legal positions and contradictory legal positions, on p 110 – the main problem with the chapter lies in the framework's excessive rigidity. Though Halpin's discussion is not unilluminating, it tries to cram rights-discourse into a structure of three categories: abstract, general concrete, and particular concrete. Because rights-discourse pertains to norms of widely varying degrees of abstraction, and because Halpin's model effectively disregards all intermediate levels of abstraction, the model is far too cramped. (His rejection of a multi-layered approach to rights-discourse, on p 157, is unconvincing.)

Even more problematic is the sixth chapter, which applies Halpin's framework to judicial opinions and scholarly writings. In the course of attacking H.L.A. Hart,

 © The Modern Law Review Limited 1999

Lord Devlin, and David Feldman (among other targets), Halpin repeatedly points out an absence of logical entailment between the premises and the conclusions of the arguments which he is examining. Unfortunately, apart from seizing on a few incautious sentences in Lord Atkin's *Donoghue* v *Stevenson* judgment, Halpin does not offer any grounds for thinking that the criticised arguments have been put forward as logical demonstrations. Hart, for example, was drawing on the ideas of Ronald Dworkin, who maintains that legal principles are connected to the law's concrete materials through ties of fit and of moral worthiness rather than of logical entailment. In a somewhat similar vein, Devlin, like most judges, was undoubtedly trying to show that his conclusions were reasonable elaborations of the existing legal norms – rather than strictly logical deductions therefrom. (Assessing judicial opinions by the rigorous standards of philosophical argumentation is the legal theorist's version of shooting fish in a bathtub.) Equally, the 'floodgates' arguments by civil-liberties scholars such as Feldman against the regulation of speech are concerned with political-psychological probabilities rather than with logical necessities.

Chapter 7 makes a number of good points against Unger's highly unsatisfactory theory of rights. The chapter loses some of its force, however, by setting an impossible task for Unger – a task which, unsurprisingly, he does not fulfill. Halpin criticises Unger for advocating some supposedly novel sorts of rights that turn out to be no different from traditional rights; but Halpin arrives at his verdict by refusing to let anything count as a right (within Unger's model of society) unless it bears the characteristics of traditional rights. To some extent, then, Halpin is talking past Unger rather than squarely engaging with him.

Chapter 8 discusses the distinction between deontological and consequentialist moral theories in a sustainedly thought-provoking manner. Halpin's efforts to minimise that distinction, however, do not really take account of the willingness of some deontological theorists to accept that – in exceedingly improbable circumstances – a course of conduct can be right even though it does no good whatsoever (aside from the good of upholding the standards of rightness). Chapter 9 is a piquant exploration of the dispute between utilitarians and rights theorists. Nonetheless, it suffers from two principal shortcomings. First, it mischaracterises the fundamental axiom of utilitarianism as individualistic rather than as egalitarian and collectivist. (Under that axiom, the processes of determining appropriate public policies should take into account each person's preferences on a par with everyone else's; and, as a result of such processes, any particular person may find that his or her preferences go entirely unfulfilled.) The gap perceived by Halpin between the individualistic premises and the non-individualistic conclusions of utilitarianism will have vanished when the premises are appropriately identified. Another major shortcoming of chapter 9 is that, like chapter 7, it question-beggingly gauges its target by reference to criteria that must inevitably be unsatisfied. Those criteria for the existence of a moral theory, which are laid down in the eighth chapter, are construed by chapter 9 in a deeply anti-utilitarian way. Hence, one can hardly be surprised that utilitarianism turns out to be inadequate as a moral theory, when assessed by Halpin's criteria.

In the ninth chapter's discussion of Robert Nozick and in the concluding chapter's general summary, the underlying source of one of the prominent misconceptions in chapter 2 becomes clear. When attacking Hohfeld, Halpin maintains that a Hohfeldian liberty is reducible to the rights which protect against interference with the actions that are covered by the liberty. Anyone familiar with Hohfeld's work should instantly recognise the untenability of Halpin's position

© The Modern Law Review Limited 1999

here. A liberty to ϕ is the absence of a duty to abstain from ϕ-ing, rather than a right or set of rights against interference with ϕ-ing. Halpin is correct to observe that every such liberty in any legal regime is accompanied by some rights which protect against interference to a greater or lesser degree; but his observation derives simply from the fact that a legal regime is unlike a Hobbesian state of nature, where each liberty would be wholly unaccompanied by rights. A recognition of that fact goes no way towards undoing the distinction between a liberty to ϕ and a right against interference with ϕ-ing.

Halpin's error is so stark as to be initially baffling. Why would a plainly intelligent theorist blunder so badly? The answer emerges in the final two chapters of his book, where his argument reveals that he has conflated Hohfeld's position on legal liberties with a quite distinct position. Halpin aptly deplores the tendency of libertarian political theorists to disregard or play down the unfreedom that ensues from the duties that are correlative with individual rights. Libertarians stress the importance of rights in protecting people's abilities to act freely, but they often neglect the restrictions which rights entail – restrictions in the form of duties, which cabin freedom. (Halpin's criticism of libertarianism on this point is not novel. Similar complaints, not cited by Halpin, have been made by G. A. Cohen, Jeffrey Reiman, Cheyney Ryan, and the present reviewer, for example. Equally pertinent is the critique of the public/private dichotomy mounted by some American Legal Realists during the 1920s and 1930s – a critique importantly inspired by Hohfeld's work.) However, *pace* Halpin, Hohfeld's distinction between liberties and rights is markedly different from the libertarian stance. Hohfeld simply pointed out that the absence of a duty to abstain from ϕ-ing is conceptually distinct from the presence of a right against interference with ϕ-ing. To submit as much is perfectly consistent with submitting that a liberty-to-ϕ unaccompanied by *any* rights against interference with ϕ-ing would be hollow indeed in just about every conceivable situation involving large numbers of people. In any viable society, each person's liberty to ϕ will indeed be accompanied by rights that offer partial or full protection against interference with ϕ-ing. Those rights, as Hohfeld emphasised, entail freedom-circumscribing duties; hence, the protections which render anyone's liberties valuable must always consist in the curtailment of others' liberties. In short, far from being a forebear of the libertarians, Hohfeld – with his emphasis on the correlativity of rights and duties – is one of the great forebears of the critics who expose the inadequacies of libertarianism.

Finally, one minor slip should be corrected. In Chapter 6, Halpin mentions that 'Learned Hand ... was on the Supreme Court that decided *Brown*' (p 171 n148). In fact, for better or for worse, Hand never became a Supreme Court Justice.

*Matthew H. Kramer**

Reziya Harrison, **Good Faith in Sales**, London: Sweet and Maxwell, 1997, xliii + 819 pp, hb £78.00.

Having long neglected the topic of good faith, English contract lawyers are now endeavouring to make up for lost time. To some extent, this recent surge of interest has been prompted by the good faith element of the general test of fairness in the Directive (and implementing Regulations) on Unfair Terms in Consumer

* Churchill College, Cambridge.

 © The Modern Law Review Limited 1999

Contracts. However, questions of good faith are not restricted to unfair terms, nor indeed to consumer contracts – they are quite simply pervasive. It is with some anticipation, therefore, that we turn to Reziya Harrison's book, trailed in Lord Bingham's Foreword as dealing with 'the most important contractual issue of our time' (p vi).

The book is divided into six unequal parts. Part One (a mere 30 pages) advances the provocative thesis that, contrary to popular belief, English law has long recognised the principle of good faith in contracts. Parts Two and Three (nearly 500 pages) discuss the good faith in duty of candour and accuracy in both pre-contractual and post-contractual dealings, as well as (in the case of the latter) the duty of loyalty. Part Four (some 150 pages) is devoted to remedies (particularly rescission), with the duty of loyalty operating as the key good faith principle. Part Five, like Part One, comprises only a couple of short chapters, here commenting on the Unfair Terms in Consumer Contracts Regulations, sketching the good faith provisions found in some other jurisdictions, and arguing that we should not 'resist the rediscovery of good faith law in this country' (p 685). Finally, Part Six brings together a selection of materials, drawing on the Directive on Unfair Terms in Consumer Contracts, the Investment Services Directive, the Misrepresentation Act 1967, the Unfair Contract Terms Act 1977, the Sale of Goods Act 1979, and the Financial Services Act 1986; and, the book rounds off with some short thoughts on the influence of Aristotle, Aquinas, Grotius and Pothier on good faith in English law.

Before we dip into this very large book, however, we should follow the author's example by clarifying our terminology. Contract writers routinely distinguish between good faith in a 'subjective sense' (meaning honesty in fact, or a clear conscience) and good faith in an 'objective sense' (meaning compliance with standards of fair dealing). Without resorting to this terminology, Harrison declares that her concern is with good faith in the sense of fair dealing and acceptable conduct relative to the parties' reasonable expectations; that '[t]he legal opposite of good faith is not bad faith; it is a breach of good faith, or a lack of good faith' (p vii) – in other words, *not* acting in accordance with the (objective) rules of fair dealing; and that '"[b]ad faith" normally means a deliberate misrepresentation or a deliberate misuse of a power' (ibid), that is, breach of good faith in the subjective sense.

It is Parts One and Five of the book that catch the eye, the former, as we have said, boldly challenging the commonplace that English law has no principle of good faith in contracts. Contrary to the conventional wisdom, Harrison's thesis is that '[t]here is to be found in English case law, more particularly in the late eighteenth and nineteenth century case law, a reasonably consistent set of fair dealing principles, or implied contractual terms of fair dealing, as incidents of certain types of contracts' (p 4). Granted, there was 'a reaction against good faith ... roughly from the late 1870's [sic] onwards' (p 7) when a narrow literalism took hold; but the author's claim is that English law has a substantial, if suppressed, jurisprudence of 'fair dealing principles ... based on a doctrine of objectively reasonable expectations' (p 18). In this light, Harrison remarks, the new Regulations on Unfair Terms in Consumer Contracts might be regarded as a case of 'carrying coals to Newcastle' (pp 11–12).

Harrison takes on the opposition in Part Five. First, she confronts the Euro-phobes. Even though good faith nowadays might seem to be inspired by European single-marketism, it is an idea with an historic English pedigree; and '[g]ood faith/ fair dealing principles are in themselves quite simple' (p 685). Secondly, to those

© The Modern Law Review Limited 1999

who think that general provision for fair dealing should be left to Parliament, Harrison suggests that judicial re-imposition of principles of fair dealing would probably be welcomed at Westminster and might indirectly have 'a litigation-quieting tendency' (p 686). Thirdly, short shrift is given to the Jeremiahs who fear that adoption of good faith would signal the death of contract as we know it. Although Harrison seemingly wants to assuage such concerns, she might have taken the opposite approach, conceding that it would indeed be the end of contract as we know it but arguing that contract would be all the better for it. Finally, there is the objection that 'the re-introduction of good faith would be bad for commercial life' (ibid), to which Harrison offers a dual response – namely, that commerce did not in fact seem to suffer during the period when (on the book's historical thesis) good faith was openly accepted in the English courts and that it does better with a few simple principles of good faith dealing than with a weight of detailed regulation. Harrison might also have said that, if good faith is bad for commerce, it is strange that it should be embraced by both the Lando Commission in its *Principles of European Contract Law* and by UNIDROIT in its *Principles of International Commercial Contracts*; that there is evidence that good faith facilitates trust, confidence and co-operation in commerce (see, eg, Deakin and Michie (eds), *Contracts, Co-operation, and Competition* (Oxford: Oxford University Press, 1997)); and that respect for the parties' reasonable expectations enhances rather than diminishes commercial calculability.

Turning to the central core of the book, Part Two opens by considering fair dealing principles during the formation phase of contracting. Harrison's survey covers contracts for the sale of personal property (chapter 3), insurance and surety (chapter 4), the sale of land (chapter 5) (note here pp 110–115 for a key analysis of various states of mind accompanying misrepresentations, but beware para 5.12 where there surely are typographical errors in the starring system employed to mark out particular degrees of blameworthiness), shorter leases (chapter 6) and company shares (chapter 7). In general, the expectation is that parties will act with due diligence in the particular circumstances but that, where one party must rely on the other, the latter will respect the former's dependency by making accurate and full disclosure of material facts. This leads to a consideration of questions concerning knowledge and notice (chapter 8) (which are central to whether, for fair dealing purposes, there is judged to be an asymmetry of information between the parties), sales involving vulnerable parties (whether the vulnerability is, as Harrison puts it, 'spontaneous' or 'induced') (chapter 9), and exclusionary terms (chapter 10). Readers will find plenty of food for thought in these chapters, not least in the author's consideration of whether good faith may be excluded, her position being that standard expectations may be displaced provided that exclusion is effected fairly. In other words, the master principle of fair dealing provides not only for the default rules but also for modifications to those rules.

The focus in Part Three moves to good faith in the performance of contracts, where academic contract lawyers have become accustomed to talking about an implied term for co-operation as the governing idea. Here, Harrison elaborates what she calls the good faith duty of loyalty (see p 445), the relevant expectation being that a party should not act in ways that are inconsistent with the paramount purpose of the contract (see p 448), and the paradigm of which is the rule against derogation from grant (chapter 11). The principle is then traced (in chapters 12 and 13) through vendors' rescission clauses, purchasers' 'subject to approval of title' clauses, termination of licence provisions, and various kinds of approval and consent clauses, and the like.

 © The Modern Law Review Limited 1999

In Part Four, the principle of loyalty is carried through into a discussion of remedies, opening (in chapter 14) with a powerful critique of the modern law on penalties, deposits and forfeitures. From a good faith perspective, as Harrison says, there 'seems no good reason why the parties should not agree a large price for a particular broken promise which is seen as important' (p 546); however, no court should permit such a provision to be fully liquidated, rather than simply cashed as security, unless it is satisfied that there has been fair dealing between the parties and that enforcement is in line with their reasonable expectations. In the development of the modern law, Lord Halsbury's well-known judgment in the *Clydebank Engineering and Shipbuilding* case is singled out as derailing the earlier appreciation of penal provisions, the author suggesting that, if one could 'give decent burial to Lord Halsbury's heresy, and the problems it generated, the time may be thought to be ripe for the emergence of a unified doctrine [concerning all species of "penal" provisions]' (p 556). In this light, it would have been interesting indeed to have had the author's thoughts on Lord Hoffmann's speech in *Union Eagle* v *Golden Achievement* (noted at 61 MLR 255). Part Four continues with two chapters (15 and 16) on rescission, the latter unravelling the confusion concerning the availability of this remedy following the execution of contracts for the sale of land; and it concludes with a somewhat compressed chapter (17) on good faith and the Misrepresentation Act 1967 (in which the commentary on the measure of damages under section 2(2) is not assisted by a poor page layout at pp 676–677, a typographical slip at p 677 when section 1(1) incorrectly stands for section 2(1), a cryptic reference at the same page to 'the McGregor view', and some throwaway references at p 678 to the 'warranty cap' concept in the *South Australia Asset Management* case (noted at 61 MLR 68)).

What, then, are we to make of this work, in which the author conspicuously sets her own agenda, adopts her own conventions, and makes no secret of her passion for, and fascination with, the case-law of the earlier good faith period? On the credit side, the author's enthusiasm generates a commentary that is engaging, lively, and refreshingly no-nonsense and that, at the same time, retrieves a once-flourishing and valuable jurisprudence of fair dealing. On the debit side, however, because the author's particular line of inquiry can be, and to a considerable extent is, pursued at some distance from much of the modern literature on good faith, the book makes only a limited contribution to contemporary good faith debates. It is not so much that landmark C20 commercial sales cases such as *Arcos* v *Ronaasen* and the *Hansa Nord*, which one might have expected to feature prominently in the discussion, are not cited or considered; rather, it is that Harrison's uncovering of the good faith principles of yesteryear does not always speak directly to today's questions. So, for example, those who conceive of good faith as context-contingent principles of fair dealing (cf Lord Steyn's important discussion at (1997) 113 LQR 433) might have welcomed more analysis of the way in which reasonable expectations are to be layered and integrated – drawing on formal legal doctrine, the informal norms of trading communities, and the particular signals and circumstances attending any one specific transaction. Equally, those who advocate non-contingent (imposed) principles of fair dealing in at least some transactional settings (eg, where one party is 'vulnerable') might feel that Harrison could have said more about the protective dimension of good faith (and about whether reasonable expectations are prior to, or based upon, such imposed standards).

Such caveats, however, strike altogether the wrong concluding note. This is a book in which the author combines both a scholarly pleasure and a practical purpose in presenting the fruits of her research for the stimulation, provocation and

© The Modern Law Review Limited 1999

assistance of the legal community. If the present rush of interest in good faith is not sustained, it will not be for want of Harrison's considerable efforts.

*Roger Brownsword**

Tony Weir, **Economic Torts**, Oxford: Clarendon Press, 1997, xiii + 130pp, hb £17.99.

Weir's lively, entertaining and stimulating Clarendon Law Lectures of 1996 now appear, slightly revised, in book form. No staid academic treatment this! Weir writes in the genre that he has made his own – the legal polemic – splendidly disdainful of conventional proprieties. In 77 pages of text (and Weir has the audacity (p 1) to describe Heydon's *Economic Torts* as a 'little book', albeit 'excellent'!) he eschews comprehensive exposition of the law and instead picks a meandering path through a number of controversial aspects of the economic torts. Whether advancing criticisms of the existing law, or defending it against what he considers to be unjustified attacks, he proceeds with sparkling verve, lacing his account with numerous digressions on this, that or the other (from a disquisition on *The Pickwick Papers* to a short biography of the unfairly neglected Mr Justice Wills, via reminiscences about teaching in Texas either side of lunch at the Black-Eye Pea). His analysis of the English law is supplemented by useful comparisons made with the law in Germany, France and the United States, the essential features of which Weir sets out clearly and concisely. In a lengthy appendix, he includes full transcripts of a number of decisions that will not be readily available to many readers.

There is much to relish in Weir's witty and elegant prose. He tells a good story, be it of how the woodmen 'got the chop' in *Allen* v *Flood*, or of the American legal battle between Tuttle the barber and Buck the banker – 'mega-Buck'. (The barber had a close shave as a result of some cut-throat competition, but the court thought that his case was cut and dried). Yet, for all the humorous asides, Weir often gives the impression of writing at the end of his tether. Decisions are not just 'wrong' but 'wrong-headed', 'scandalous' or 'disgraceful'; developments in the law are 'wretched'; the English legal system embodies 'objectionable values'. It's thrilling stuff, but one looks in places for a little more explanation to go with the excoriation.

I cannot hope to emulate Weir's flamboyant style, but I hope that I have grasped his argument well enough – in spite of its many twists and turns – to summarise it here. Weir argues that English law has erred in treating 'interference with the plaintiff's rights' as a principle of liability for the intentional infliction of economic loss; in his view, a rival principle – that of liability for 'intentionally causing harm by wrongful means' – provides the only adequate foundation for the economic torts. This error, he submits, has lead to a number of unfortunate developments in the modern law. First, the focus upon the plaintiff's rights rather than the defendant's wrong meant that 'it might seem that the tort applied to all rights, whatever their source' (p 31). This is not only misleading (eg because violation of rights under a trust is actionable in equity and not in tort) but also dangerous, in that it might encourage the extension of liability to violations of rights of a purely moral or social character. Secondly, neglect of the defendant's wrong has lead some judges to express the view that bare interference in another's contract, without any breach, might give rise to liability in the absence of wrongful conduct

* University of Sheffield.

 © The Modern Law Review Limited 1999

on the defendant's part (see especially *Torquay Hotel Co Ltd* v *Cousins* [1969] 2 Ch 106 at 137, per Lord Denning). Thirdly, the word 'interference' obscures a fundamental distinction between *persuading* another not to perform a contract and *preventing* him from so doing: the former is wrongful in itself, but the latter is not and ought not to give rise to liability unless some independently unlawful act is involved (as where the defendant steals the only tools available to do the job). Lastly, the extension of liability to cases of prevention without persuasion has created the risk that actions might be brought by those who were not the intended target of the defendant's action but merely the 'by-blow' victim. (Where persuasion is involved, it can at least be assumed that the defendant is out to get, or get something from, the plaintiff).

It seems to me that Weir makes a lot of fuss about not very much. Most of his criticisms focus upon developments that might have happened but have not. Procuring a breach of trust continues to be governed by equity, not the law of tort (*Metall and Rohstoff AG* v *Donaldson Lufkin & Jenrette Inc* [1990] 1 QB 391), and the courts have so far resisted attempts to extend liability to the violation of rights of a purely economic or social character. (As Weir notes, the Court of Appeal has recently rejected an argument that interference with 'parental rights' should be actionable in tort: *F* v *Wirral Metropolitan Borough Council* [1991] Fam 69.) Additionally, Weir is far from alone in attacking Lord Denning's proposal that bare interference with a contract should give rise to liability, and it seems in any case that no subsequent decision gives it unequivocal support. (Even Lord Denning's own examples – eg the poisoning of the opera singer contracted to perform for the plaintiff – are cases where liability might be justified on the basis that the plaintiff was injured by unlawful means). And, the odd throw-away remark aside, there appears to be no judicial enthusiasm for the imposition of liability in respect of otherwise lawful conduct (not involving any element of persuasion) which prevents another from performing a contract.

There is more to be said about Weir's suggestion that the defendant's act must be 'directed at' the plaintiff if it is to give rise to liability. By this, I take him to mean that it is insufficient that the defendant should *know* that his wrongful act will harm the plaintiff: he must actually *want* to cause him harm (or to benefit at his expense). The test is one of *direct* intent, not *oblique* intent. Weir's concern is that, without such a requirement, there would be a large increase in the number of plaintiffs, 'something', he adds, 'we normally try to avoid' (p 38). I am sympathetic to this argument, but think that it needs more supporting analysis if it is to persuade the contrary-minded. Weir, however, seems to treat the matter as self-evident, and restricts himself to heaping ridicule upon a couple of decisions which he regards as misguided. I am not sure that his examples are well-chosen. Consider his treatment of the Court of Appeal's decision in *Millar* v *Bassey* [1994] EMLR 44. In that case, the defendant pop star, Shirley Bassey, withdrew from a recording project in breach of contract with her production company and was sued by musicians who had been engaged for the sessions by the company; the court declined to strike out the actions, but noted that Bassey might be able to raise a defence of justification. 'Dear me!', exclaims Weir, '[i]t is easy to see how wrong this decision is' (p 19). Well, not to me it ain't. Nor, I guess, would it have been clear to the musicians. After all, they might well have found themselves deprived for a time of their livelihood, having turned down alternative engagements in order to play with the talented but temperamental Ms Bassey. Maybe that is just one of the risks that you take as a session musician. Maybe – but it is not as blindingly obvious as Weir makes it appear.

© The Modern Law Review Limited 1999

A better target for Weir's censure would have been the much-criticised County Court decision in *Falconer* v *ASLEF and NUR* [1986] IRLR 331. In that case, a British Rail passenger succeeded in recovering damages from the unions in respect of expenses he incurred when industrial action prevented him using his return rail ticket on the day he had intended. The decision is more noteworthy than that in *Miller*, where the plaintiffs were at least involved with the defendant in a specific business venture, in which they were to perform a necessary role; in *Falconer*, there was no link of this sort at all. *Falconer* makes clear how widely the net of liability might be cast if the 'directed at' requirement advocated by Weir should be rejected. It also makes it easier to understand the adverse consequences that would follow from this, eg that industrial disputes would no longer be left for resolution by the participants themselves but would be susceptible to interference from third parties with no interest in a peaceful and lasting settlement. This concern did not, of course, prevent the last government from introducing a statutory cause of action which extends to *anyone* who suffers a disruption in the supply to him of goods or services arising out of unlawful industrial action (Trade Union and Labour Relations Consolidation Act 1992, s 235A). Yet to be repealed, this provision serves to emphasise that the question of by-blow victims has no self-evident answer and that mere assertion is unlikely to turn the debate in one's favour.

In addition to the specific matters already noted, it is worth moving to a higher level of abstraction to remark upon Weir's view of the economic torts as an essentially unitary area of law, resting upon a single principle of liability for the intentional and wrongful causing of harm. While I am sympathetic to the attempt to reduce complexity and anomaly by identifying general principles underlying the law of tort – which is in a terrible mess and needs tidying up – I fear that Weir's analysis is excessively reductive. I would subscribe to the view that the economic torts are founded upon *two* separate and distinct underlying principles, one imposing liability for procuring the violation of a right (more specifically, violation of a private right actionable by the injured party), the other imposing liability for the intentional and unlawful interference with economic interests. This bipartite scheme is well-supported by the authorities (see, eg, *Associated British Ports* v *TGWU* [1989] 1 WLR 939, Court of Appeal) and seems correct in principle. In cases of violation of right, the defendant participates in the wrong committed by the party who actually injures the plaintiff and against whom the plaintiff has a prima facie cause of action; the defendant's liability is secondary to that of the primary wrongdoer. By contrast, in cases of unlawful interference, the liability does not rest upon the defendant's participation in a third party's wrongful conduct but simply upon the defendant's own use of unlawful means; the liability is not secondary but primary in the sense that it is independent of the liability (if any) of a third party. In such cases, the question arises why the unlawful conduct should be actionable at the suit of a plaintiff who may well have suffered no violation of his (pre-existing) rights. It may be that the law is justified in imposing restrictions upon actionability here which are inappropriate in the case of a violation of rights lest liability be spread too wide (and lest existing restrictions upon actionability, eg for breach of statutory duty, be undermined).

Distinguishing these two principles is necessary for any further attempt to develop general principles of secondary civil liability (currently conspicuous by their absence in English law) by drawing together principles of dishonest assistance in equity, joint tortious liability and procuring the violation of rights. Weir is sceptical about such attempts, but advances no more concrete reason

 © The Modern Law Review Limited 1999

against them than that '[a]ccessory liability in the criminal law has not been joyous, nor has it in equity' (p 32n). I would submit that it is precisely the failure to see interconnections across legal categories that accounts for the dolorous expression on the faces of tort and equity lawyers when looking at secondary forms of liability.

Time to take stock. Weir's account of the economic torts provides plenty of entertainment and plenty of food for thought. But most of the ideas presented here are set out, more clearly and concisely, elsewhere in Weir's published oeuvre (eg in the note sections of his casebook). I cannot see this book, which costs nearly £20 for only 77 pages of original text, having a wide readership. Perhaps the publishers should reconsider their strategy for future Clarendon Lectures.

*Ken Oliphant**

Tony Prosser, **Law and the Regulators**, Oxford: Clarendon Press; 1997, xx + 313 pp, hb £40.00.

The interface between law and economics is at its most visible in the field of regulation. *Prima facie*, the two disciplines appear to co-exist relatively happily, concerned with different aspects of the regulatory process and performing complementary functions. In reality, however, they make uneasy bedfellows, representing competing paradigms for the organisation and understanding of regulatory systems: law prioritising issues of fairness; economics, questions of efficiency. In the United States, where the constitution's pervasive presence gives legal standards of legitimacy much greater prominence and authority than they have in the United Kingdom, the fairness paradigm has dominated for much of the twentieth century, arguably at considerable cost to the efficiency of regulated industries. Here, by contrast, lawyers and legal concerns have been very much the poor relation, deliberately marginalised in the design of the new wave of regulatory institutions which accompanied the Conservative governments' privatisation programme, in a explicit reaction against American experience. Although we may feel a degree of *schadenfreude* at the subsequent controversy which has surrounded these new regimes – raising, as it does, the very questions of fairness, accountability and legitimacy which are central concerns of public law – lawyers' impact on academic and political debates on these issues has been equally marginal. Indeed, as a legal academic who is interested in regulation, one is often left with the unflattering impression that one is viewed by economists and policy-makers as a mere technician with little to say on larger questions.

Against this background, Prosser's book, *Law and the Regulators*, is doubly important. First, it is a major addition to the limited legal literature on regulation and considerably improves the accessibility of the subject to those with little grasp of economics. Secondly, and more significantly, it presents a direct challenge to some of the assumptions underpinning the dominant economic paradigm. It does so both in terms of their descriptive accuracy, insofar as they purport to explain the nature and purposes of existing regulatory regimes, and by calling into question the normative assumption that an economic rationale is not only a *necessary* but also a *sufficient* condition for regulatory legitimacy. One of Prosser's aims, therefore, is to demonstrate the complexity and pluralism of UK regulation, as regards the range

* School of Law, King's College London.

© The Modern Law Review Limited 1999

of tasks undertaken, as well as the variety of institutional forms which are employed. In so doing, however, he also addresses the question of how, in a democracy, regulators can justify exercising power with no direct democratic mandate; in other words, how can regulation be made legitimate? His claim is that both procedural and substantive justifications can be found for this which permit regulators to take decisions based on a variety of rationales, not merely confined to mimicking the effects of a competitive market. Thus, an important theme running throughout the book is the attempt to defend social rationales for regulation against the charge of arbitrariness.

The first two chapters of the book deal with general issues relating to the purposes of regulation and the development of regulatory institutions in the United Kingdom, setting the scene for later chapters on individual regulated industries. In Chapter One, Prosser locates much of the recent controversy over the role and accountability of regulators in failure – as previously during nationalisation – to address the fundamental questions of why regulation takes place and what basic principles should guide decision-making. He identifies three distinct regulatory tasks – regulation of monopoly, regulation for competition and social regulation – which rest on different theoretical foundations and which may have contradictory policy implications. Although it is the first two of these goals which has been emphasised in official justifications for regulation and in subsequent debates, UK regulators are in fact *legally* obliged to carry out all three. Indeed, drawing upon pre-existing common law principles and developing European doctrines, as well as specific regulatory statutes, Prosser claims to find a consistent legal principle enforcing universal access to basic services, resting on egalitarian or rights-based, rather than economic, arguments. Nevertheless, in practice, the regulators are left with substantial discretion as to how to fulfil their legal mandates and how to reconcile the conflicts between different tasks. This makes it essential to address issues of procedures and accountability as well as the substance of decision-making. Here, too, Chapter Two demonstrates that there has been an almost wilful refusal to learn the lessons of earlier regulatory experience, not only from the United States, but also from pre-nationalisation regulatory regimes in the United Kingdom. Whilst accepting many of the criticisms levelled at these earlier institutional forms, Prosser views it as unduly simplistic to ascribe the problems they encountered wholly, or even mainly, to excessive legalism. Clearly, an *ad hoc* rather than a principled approach to institutional design does not remove the fundamental difficulties in reconciling independent regulation with accountability and legitimacy – illustrating the truth of the observation that those who ignore history are destined to repeat it.

The next seven chapters – examining regulatory goals and institutions in the four utility industries (telecommunications, gas, water and electricity) as well as in the transport (bus and rail), aviation and independent broadcasting sectors – form the bulk of the book. Once again, Prosser's aims are primarily descriptive, though he goes well beyond merely outlining statutory objectives and procedures, focusing instead on the way in which the regimes have operated in practice and how they have developed over time. However, his analysis of the strengths and weaknesses of the individual regimes also forms the basis for the final chapter, which returns to a general consideration of regulatory legitimacy. Despite being critical of numerous aspects of current regulatory practice, Prosser's conclusion is that independent, industry-specific regulation is here to stay, having considerable advantages over the alternatives of a return to nationalisation, no or self-regulation, or reliance on the general competition authorities. Ironically, not the least of these

 © The Modern Law Review Limited 1999

advantages, as far as the utility regulators are concerned, is that the new regimes have enabled the development for the first time of an explicit scheme of social protection for consumers in respect of basic goods and services. Nevertheless, Prosser argues that we can learn from overseas experience in order to guide future developments in regulation. As one might expect from a member of the Sheffield diaspora, he draws on American proceduralism as a model for improving regulatory openness and for accommodating the plurality of interests in regulatory decision-making. However, rejecting the idea that procedures alone are sufficient to legitimate decision-making, he also recommends the adoption of something like the respective French and Italian concepts of *service public* and *servizio pubblico*. Such a move would, he argues, provide an accepted structure within which debates about regulation could take place, as well as a body of principle to underpin social regulation equivalent to the body of economic theory which guides economic regulation. Far from being alien impositions on the British system of regulation, he demonstrates that both sets of recommendations would in fact build on developments which have already taken place in British regulation, albeit in a piecemeal and informal fashion, or would coincide with reform proposals made by other commentators, which he discusses in the final section of the chapter.

As someone who has struggled to understand and keep up with developments in the regulation of just one industry (electricity), one of the things that impressed me most about *Law and the Regulators* was Prosser's mastery of so many disparate and complex regimes. Admittedly, the price of breadth of coverage is that there is probably little in each of the substantive chapters that is new to anyone who is already familiar with the sector in question – though the book provides an excellent introduction for those who are not. However, the real advantage of the comparative approach is that it allows Prosser to make realistic assessments of the merits of competing proposals for reform of utility regulation. Despite the title, it is clear that this is really a book about utility regulation and that other industries have been included for their ability to cast light on the problems facing utility regulation rather than because of any desire for comprehensiveness for its own sake. Thus, for example, experience in the bus industry constitutes an effective rebuttal to arguments that industry specific regulation is unnecessary or that it increases the risk of litigation and legal complexity. Similarly, civil aviation provides a successful model of regulation by commission and through formal hearings, as well as support for the proposal that government ministers should be obliged to give guidance to the regulators in order to structure their discretion and improve accountability.

The book is, however, less satisfactory in its discussion of the purposes of regulation. Prosser is undoubtedly correct to argue that understanding the aims of regulation is essential to any adequate consideration of the conditions of regulatory legitimacy. He is also correct to point out that, both as a matter of law and practice, regulatory objectives are rather more complex than is often assumed. The problem is, though, that he does not really address the disputes about regulatory purposes on their own terms. First, it is far from clear how much weight should be given to existing legal obligations as statements of regulatory purposes – both because of the rather haphazard way in which the statutory duties were drafted and because of general scepticism about the importance of law in shaping human activity. Thus, for example, the regulators often seem to invoke their statutory duties as *justifications* for decisions rather than as reasons. Similarly, while there is growing consensus that consumer protection ought to become a primary rather than a secondary duty for the utility regulators, few expect this to make much difference to their day-to-day priorities. Indeed, the regulators already claim that this is their overriding goal.

© The Modern Law Review Limited 1999

Of course, the precise wording of the utility statutes is potentially an important determinant of regulatory behaviour insofar as decisions may be challenged in the courts. Less obvious, however, is the relevance of the common law principles which Prosser also cites as evidence of a social rationale for regulation. Nowhere does he address the question of how these principles could be used against utility companies, nor whether they have in fact fallen into desuetude. While, in a country like New Zealand, where essential industries have been privatised with no specific regulatory oversight, there may be little choice but to revive eighteenth and nineteenth century cases, it is at least arguable that they have implicitly been overruled, or at least rendered redundant by the plethora of statutory and licence obligations imposed on British utilities.

The second problem with Prosser's handling of this issue is that he seems not to recognise that the reasons for disagreement over regulatory purposes are essentially normative in character. An economist, or a regulator, who argues that social regulation, unlike economic regulation, is inherently political and properly the task of ministers is unlikely to change her or his mind simply because it is pointed out that the law says otherwise. Their underlying assumption is that regulation *ought* to be concerned only with promoting competition and efficiency and hence that statutory obligations and procedures should be reformed so as not to interfere with these goals. It *is* possible to mount an effective critique of this technocratic attitude, but one has to proceed from a demonstration that economic regulation is in reality impossible to divorce from questions of fairness, followed by arguments as to why a market-oriented approach is inadequate in respect of basic industries, or at least an explanation of why social regulation has persisted in spite of the ideological motivation behind privatisation. Prosser does none of these things. Accordingly, his recommendations for securing procedural and substantive legitimacy will convince only those who are already prepared to accept the premise that a plurality of regulatory rationales is desirable.

Ultimately, therefore, while *Law and the Regulators* is a welcome contribution to debate and contains much that is of interest, it is unlikely by itself to effect a paradigm shift in British attitudes towards regulation.

*Aileen McHarg**

Peter Goodrich (ed), **Law and the Unconscious: A Legendre Reader**, Basingstoke: Macmillan Press, 1997, xvi + 270pp, hb, £45.00

Here is a manifesto of shadows. *Law and the Unconscious* brings together for the first time a substantial collection of Pierre Legendre's writings. As a collection it is a fine representation of a work in progress: the patient and exhaustive study of the central, interlocking structures of Western law. It is a work that proceeds with an archival passion and a drive that is fuelled both by the sense that this important work has been ignored by the legal establishment, and that what is at stake is an understanding of the destiny of Western law.

The book could perhaps best be described as organised around the need to identify and describe law's inseparability from the mythic definition of social being through the creation of both the subject and the community. For Legendre, these myths are best approached through a form of psychoanalysis, a discourse which does not

* School of Law, University of Bristol.

 © The Modern Law Review Limited 1999

assume an exteriority and mastery of myth, but is rather a continuation and redefinition, a contemporary myth of the law. Psychoanalysis can offer an account of the subject's attachment to legal and political institutions through the capturing and manipulation of unconscious desire. Law is essentially poetic and fictive: it creates the subject and the possibility of civilized being. These insights are introduced and presented in a way which does not dwell upon what some have seen as the technical and abstruse developments of psychoanalytic theory.

The introduction presents Legendre's work as drawing on Freud's later texts on the psychoanalysis of social organisations and Lacan's re-definition of Freudian analysis. Lacan privileged the role of language in the creation of the unconscious, stressing the centrality of the 'mirror stage', where the human child takes on the first fictive sense of personal autonomy and coherence. One of the longer pieces in the book, 'Hermes and Institutional Structures: An Essay on Dogmatic Communication', focuses on the dynamics of the institutionalisation of speech, and the central role of law in this process. The human comes to being in the separation from the mother and speech is founded in this separation. It is this complex and necessary social function that is also described in terms of the Narcissus myth in the final essay, 'Introduction to the Theory of the Image', The social order cannot allow the human subject to fade away and die like Narcissus: social being involves mastering the essential absence on which we are founded. Society has to take up the creation of a sustaining symbolic level to maintain the subject. This is the work of the law and its fictions.

Law's work is termed 'the Text without a subject'. The Text is a principle of social reproduction implied into all social organisation and necessary to any definition of the human. It is 'without' the subject as it does not 'belong' to anyone. The State cannot be explained by reference to the interests of class or conventional theories of sovereignty. The Text is also without the subject in that power, for Legendre refers always to the absent source on which both law and politics are founded. This could perhaps be best exemplified by the Roman Emperor Justinian, the living embodiment of the divine and absent source of the law. Ultimately, like the subject itself, both the law and political power are built over the void.

In order to understand the construction of legal order, Legendre returns to the codes of divine and human law that lie at the base of the Western tradition. The key to the whole theoretical endeavour, however, is the notion of the 'revolution of the interpreters', a complex and protracted process that played itself out between the eleventh and thirteenth centuries. The revolution saw the ascendancy of the Pope as the true source of law, and the reception of the Corpus Iuris Civilis into European culture. Legendre's project traces the genealogy that exists between Roman Law and modern industrialised legal reason. Roman law provides the basic categories and structures for the development of secular forms of power. There are many aspects and complex links to be traced and teased out. The 'arts of writing and loving', as developed in the time of Abelard and the birth of scholasticism, are seen as holding a clue as to the nature of legal transmission. The love letter, and its protocols, are presented as emblematic of a form of erotic attachment to messages and communication. To respond to the lover's letter is to respond to a writing that commands and interpellates. The love letter carries the truth of a legal writing that can address and enrapture the subject; just as the lover is sworn to the beloved, the legal subject is held in fealty to the institution of law.

Legendre's fascination is also with liturgy and liturgical law as the definition of institutional language, the organisation of a 'dogmatics of tears', a psycho-spiritual complex which defines political power through the license and control of the

© The Modern Law Review Limited 1999

emotions. This is linked to a particularly bold thesis where the suppression of dance in early Christian Europe is seen as a prerequisite for the development of law. The refutation of the body and its desires were essential to the creation of the possibility of the subjects' adherence to culturally definitional structures, or what is strikingly termed the 'solemnization of the body through discourse'.

As much as modern law would like to believe that it operates through pure rationality and can be traced to a secular self-foundation in its own immanence, it is reliant on these structures for its own institutionalisation. In other words, institutional reason is not master in its own house. Any account or understanding of law which ignores its wider history and 'poetic' function is dangerously deluded. Chapter eight is central to this thesis. It is devoted to the elaboration of the place of the judge as a figure who repeats the interdiction that institutes civilized life, reiterating the bond that holds the subject to the institution. Legal judgment depends on the inclusion of the 'internal jurisdiction' of the priest into the role of the judge. Despite secularisation, the office of the judge must retain its symbolic function: ritually re-founding the principle of legal reason. This is an essential part of a wider consideration of the secular administrative state. Rather than replace the Church, the state effectively takes on its social function. Just as the Church had to maintain the distraction of the masses from the desire of the unregulated and pre-social body, the modern state stresses that the gaze of the multitude should be directed towards its own embodiment of power. The centralised institutions of the State effectively replace the sacred with sovereign norms and the rules of Roman law. The modern state needs to acknowledge its preserving mythologies.

It is the trajectory of this work, the return to the medieval and sacred sources of Western law that makes Legendre's work so engaging and could position *Law and the Unconscious* as an opening for many debates. What is communicated by this book is the sense in which Legendre is only at the beginning of an ongoing project. His erudition is offered as something to build upon and elaborate. One of the major possibilities that emerges is that of sustaining a Legendrian reading in and against the common law tradition. The book also has great relevance for the debates within postmodernism and theology, and the complex areas developing around the concerns of justice, ethics and law. *Law and the Unconscious* is an essential text for any scholarship that seeks to be informed by the best of continental thought.

*Adam Gearey**

Sally Sheldon, **Beyond Control: Medical Power and Abortion Law**, London: Pluto Press, 1997, 224 pp, hb £40.00, pb £13.99.

The legal regulation of abortion has been a highly contested and much-discussed issue in feminist work, and at first one might wonder why another book on this subject has been published – is it not likely simply to retrace a well-trodden path? Reflection, however, reveals that the vast bulk of work in this area has been undertaken by North American writers, operating in a very different social and legal context from that which exists in the United Kingdom. In the United States, abortion is a constitutional issue framed in terms of privacy rights; in Canada, the Supreme Court struck down the Criminal Code restrictions of abortion as violating women's rights under the Charter of Rights and Freedoms. Sheldon's book, with its focus on the situation in England and Wales, is a valuable contribution to the

* Department of Law, Birkbeck College, London.

 © The Modern Law Review Limited 1999

understanding of how the regulation of abortion has developed in the absence of such overriding constitutional guarantees. That the work does not extend to a discussion of the position in Northern Ireland is regrettable but understandable, since to encompass the situation there would have required discussion of many additional issues, making the book a very different proposition.

In fact, one of the strengths of this book is Sheldon's determination to confine her discussion to particular geographical, temporal and analytical boundaries. She resists the temptation to tackle the whole gamut of debates surrounding abortion, and in so doing manages to produce a carefully focused work. The book takes as its starting point the time just before the introduction and enactment of the Abortion Act 1967. It continues through subsequent developments such as the changes introduced by the Human Fertilisation and Embryology Act 1990, concluding with an examination of legal responses to the introduction of the antiprogestin abortion drug, RU486, in the 1990s. Analytically, the work is unified by its focus on a central theme, that of understanding the regulation of abortion in England and Wales as a phenomenon of medicalisation, defined as the pre-eminence of a medical discourse and the marginalisation of other understandings or knowledges. Building upon Foucault's work, Sheldon examines the recent history of abortion law as a process which shifted the primary locus of control over pregnant female bodies from the realm of the criminal law to that of the medical professions. She also draws upon insights from feminist theory and its critiques of the public-private divide. Some readers may be weary of seemingly obligatory citations of Foucault: all too often such 'name-checks' are accompanied by painfully arcane and obfuscatory prose. In this work, however, the author develops and applies the insights of Foucault's theories in a way which integrates analysis and description. The prose, furthermore, is consistently clear and free from jargon.

One frustrating aspect of this work lies in the limitations of the sources. The discussion of how the regulatory scheme operates in practice draws upon surveys of doctors and anecdotal reports from women seeking abortions. This information is patchy and rather too old to constitute a reliable account of current practice. Perhaps no better evidence is available in this area. Sheldon acknowledges the difficulty of relying upon such problematic evidence, and deals with it by carefully acknowledging its limitations. Throughout the work she notes the partial and problematic nature of the information, paradoxically giving greater strength to her argument by confining her claims to those which can reasonably be supported by the available data.

Readers familiar with Sheldon's published work will probably already have read some of this book; all save the introductory and concluding chapters are revised versions of essays which have already appeared elsewhere. They do not, however, read as disjointed pieces but as integrated parts of a unified work, and the reader's understanding of them is likely to be furthered by their being placed in the context of a broader project.

Current debate about the interface of law with reproductive technologies has tended to address issues surrounding IVF and other assisted pregnancies; it might be thought that access to termination services is no longer contentious or in need of discussion. This book is a reminder that the regulation of abortion services is still an important area of study for all those concerned with questions of professional power and control of women's bodies.

*Mary Childs**

* Faculty of Law, University of Manchester.

© The Modern Law Review Limited 1999

Giuliano Amato, **Antitrust and the Bounds of Power: The Dilemma of Liberal Democracy in the History of the Market**, Oxford: Hart Publishing, 1997, xiv + 133 pp, hb £22.50.

This book provides an enlightening account of the historical developments of antitrust law in both the US and the EC. It places the observations drawn out from historical analysis into a contemporary context and explains the divergence between US and EC antitrust policy. For a text which covers such broad areas it is relatively short, but its length belies its depth. By tracing the major developments and landmark decisions on both sides of the Atlantic, Amato brings out the central themes in an accessible way. The book will certainly be of value to those who are interested in comparisons of the two antitrust systems or the theoretical and historical bases of antitrust policy.

Part One of the book outlines the policies which have shaped the development of the US system. It charts the development of the common law principles incorporated in the Sherman and Clayton Acts, which protected 'freedom of contract' and moves towards the protection of competition. The importance of economics is reflected in the rise of the Harvard School and its concentration on the protection of competitors over competition. All of the legal developments are situated against the depression in the 1930s and the post war rise of Harvard. The emergence of the thinking of Aaron Director and the Chicago school, and its critics, is covered in rather more detail. Developments are traced through cases such as *Schwinn*, *Sylvania* and *Kodak*, all of which indicate the importance of the Chicago school's thinking but show that economic theory has not completely displaced all other thinking in the courts.

Part Two examines the development of EC controls and the very different ideals which shaped that development. The main influences here are the Ordoliberalen school in Freiburg and the secondary role of competition policy in the Treaty of Rome. The greater acceptance of both private and public power in Europe and the willingness to consider industrial and social goals when making decisions are introduced in chapters 4, 5 and 6, where the controls are examined in more detail.

The final part draws together Amato's observations and contains his thesis. He suggests that the US system was based on the protection of the Jeffersonian ideal of the small producer's right to trade, while Europe wanted a framework to control the 'monsters' created by the market. Both systems recognised the importance of democratic as well as economic efficiency. But, as time passed, the importance of the former was diluted, with economic goals becoming prevalent. This change, combined with changes in the structure of the market, is highlighted in relation to the growing difficulties of oligopolistic markets. Amato challenges the concentration on 'one-dimensional' efficiency, particularly in relation to the economic concept of 'consumer welfare'. In his view, efficiency must take 'democratic efficiency' into account.

This view of efficiency is developed into an examination of the future of the EC system and the globalisation of markets (p 113 *et seq*). The way in which industrial and social goals are included in the competition system and the complex role of the Commission are challenged. Amato argues that the antitrust rules could be freed from the other goals of the Community and that the political nature of the Commission's decision-making could be reduced by focusing on behaviour which is, according to a reformulated concept of efficiency, truly restrictive. The globalisation of markets also brings new challenges, which, Amato points out, are similar to the challenges faced at the inception of antitrust. As markets increase in

 © The Modern Law Review Limited 1999

size and move outside traditional jurisdictional boundaries, economic actors may not have true market power but may increasingly have political power. These challenges mean that a new chapter in the evolution of antitrust law may be beginning.

The book is well written and develops a number of challenging ideas in an accessible way. At times, the text can be somewhat confusing, perhaps because it is translated from Italian, but its occasional infelicities are not too distracting. Its major strength is in the depth of the research and the copious references. While the text is relatively brief, the book allows the reader to follow up areas of interest. This is particularly valuable as few academics will share Amato's depth of knowledge in both the US and EC jurisdictions. As a concise history of antitrust laws, the book is indispensable. The ideas which it develops are challenging and serve as a valuable balance to the mass of economic analysis which purports to explain antitrust.

*Angus MacCulloch**

Hans Zeisel and David Kaye, **Prove It with Figures: Empirical Methods in Law and Litigation**, New York: Springer, 1997, xxii + 353 pp, hb £43.00.

The methodologies of the social and statistical sciences are frequently relevant to litigation. When they are, lawyers (often among the most innumerate of professionals) cannot afford not to understand them. Unfortunately, the language of these disciplines – which includes concepts such as sampling, P-values and statistical significance – is forbidding and frequently open to misinterpretation. In *Prove it With Figures* David Kaye and the late Hans Zeisel aim to make these concepts accessible to those with no previous experience of them. By the end of the book the reader will have been introduced to most of the fields in which statistical and empirical techniques are likely to crop up in the course of litigation: epidemiology, hypothesis testing, sampling, jury selection and DNA evidence, among others.

This is not the only book in the 'statistics for lawyers' vein. What marks it out from other such texts is its simplicity and generality. *Prove it With Figures* does not attempt to teach the reader how to use statistical and empirical tools, rather, it aims to familiarise her with the basic concepts, enabling her to understand the strengths and limitations of various techniques and to ask the right questions of their proponents. Another of the book's strong points is that the authors are sufficiently knowledgeable about both law and statistics to avoid – and to warn the reader against – various common misinterpretations. Thus they ably explain just what P-values are and how they differ from legal standards of proof. The authors are also not so beguiled by statistics that they accept statistical conventions without question: for example, they are rightly circumspect about the convention of dismissing studies that do not produce 'statistically significant' results. In short, academics, students and practitioners who want a basic introduction to the statistical and social sciences will be amply rewarded by reading this book.

*Mike Redmayne***

* Faculty of Law, University of Manchester.
** Law Department, Brunel University.

© The Modern Law Review Limited 1999

Michael Cavadino, **The Law of Gravity**, Sheffield: Sheffield University Joint Unit for Social Services Research, 1997, iv + 92pp, pb £9.50.

This short monograph summarises a number of interesting and important issues about crime seriousness. The first three Chapters discuss the meaning of seriousness, its proper role in sentencing, and address some topical issues such as the limits to ordinal and cardinal proportionality, and the relevance of prior record. Cavadino accepts that crime seriousness should set the 'upper limit and the usual range' for punishment. He rejects strict proportionality between offence and penalty, however, arguing that the pursuit of other goals might well justify a lenient sentence — such as an order for reparation, rather than custody. What is missing from his account, it seems to me, is an assessment of the extent to which non-custodial sanctions can be accommodated within a desert framework. Consider two otherwise identical cases where offender A is able to make reparation (by money or by practical help) to their victim, but offender B lacks the assets or practical skill. Or where offender A's victim is prepared to take part in a reparation scheme but offender B's victim is not. If we accept, as Cavadino does, that there is 'a certain virtue in consistency and treating like cases alike' (p 26), it must be unjust to sentence A differently from B, *unless* the onerousness of the reparation which A must undertake is broadly comparable with the onerousness of the custodial sentence which B receives.

The material in Chapter 4 (dealing with the views of practitioners on offence seriousness) is of much interest, although it has been published elsewhere in article form. It seems that social workers would make the softest sentencers (no surprises there) but it is magistrates' clerks, rather than the police, prison service employees, or the magistrates themselves, who would make the toughest. Chapter 5 (on pre-sentence reports) also contains material published before, but this is valuable research on the practical impact of the Criminal Justice Act 1991, which should be widely known. The researchers found that the Act improved the understanding between the probation service and courts, and prompted a better flow of information to report writers from the CPS. One or two minor quibbles about the book: On p 32, the author argues that since 'intention to cause grievous bodily harm' is sufficient mens rea for murder, murder is no more than 'common assault with unfortunate consequences'. This is clearly wrong, despite the citation of Rumpole of the Bailey as authority. There is an unfortunate misprint on pp 57–58, where a whole section of text has been printed twice. It is misleading (at least without further explanation) to refer to compensation orders as a form of punishment (p 25).

This is a useful little guide on seriousness of crime, which would certainly be of interest to criminal justice practitioners and to students on criminal justice courses.

*Martin Wasik**

* Faculty of Law, Manchester University.

 © The Modern Law Review Limited 1999

Information for Contributors

Contributions to and correspondence concerning the various sections of the Review should be sent to the relevant Editor:

Articles: Hugh Collins, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Legislation and Reports. R.C. Simpson, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Case Notes: Roger Brownsword, Faculty of Law, The University, Crookesmoor Building, Conduit Road, Sheffield S10 1FL.

Books for Review, Review Articles and Reviews: Neil Duxbury, Faculty of Law, University of Manchester, Oxford Road, Manchester M13 9PL.

General Correspondence: Tim Murphy, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Section Editors will be happy to discuss drafts and proposed contributions with contributors. All contributions should be typed, double spaced and include an estimate of the number of words, including footnotes. Contributions should use footnotes which make cross-references within the text as sparingly as possible. Contributions must conform to the Review's house style; a style sheet will be provided by the Production Editor on request.

Articles do not usually exceed 12,000 words (excluding footnotes). Longer or shorter articles can be considered.

Subscription Information 1999

The Modern Law Review is published six times a year. Subscription rates are as follows:

Annual subscription for unbound parts only	
UK and Europe	£73.00
Rest of the world	£75.00
North America	US$124.00*
UK and European law teachers and law students	£21.00
Commonwealth law teachers and law students	£26.00/US$43.00*
Annual subscriptions for Bound Volume Service: unbound parts plus a bound volume	
UK and Europe	£121.00
Rest of the world	£128.00
North America	US$211.00*
UK and European law teachers and students	£59.00
Commonwealth law teachers and law students	£62.00/US$102.00*

Back Issues

Single issues of the unbound parts of the last three volumes are available from the publishers at the current price for unbound parts:

£15.00 (£15.00 for ROW, US$25.00 in North America)*

Single bound volumes	£61.00 (£63.00 for ROW, US$104.00 in North America)*
Complete sets of bound volumes	£2071.00 (US$3417.00 in North America)*

*Canadian customers/residents please add 7% GST.

Orders, sample copy requests, claims and other subscription enquiries should be addressed to: Blackwell Publishers Journals, PO Box 805, 108 Cowley Road, Oxford OX4 1FH, UK. All subscriptions are supplied on a calendar year basis (January to December).

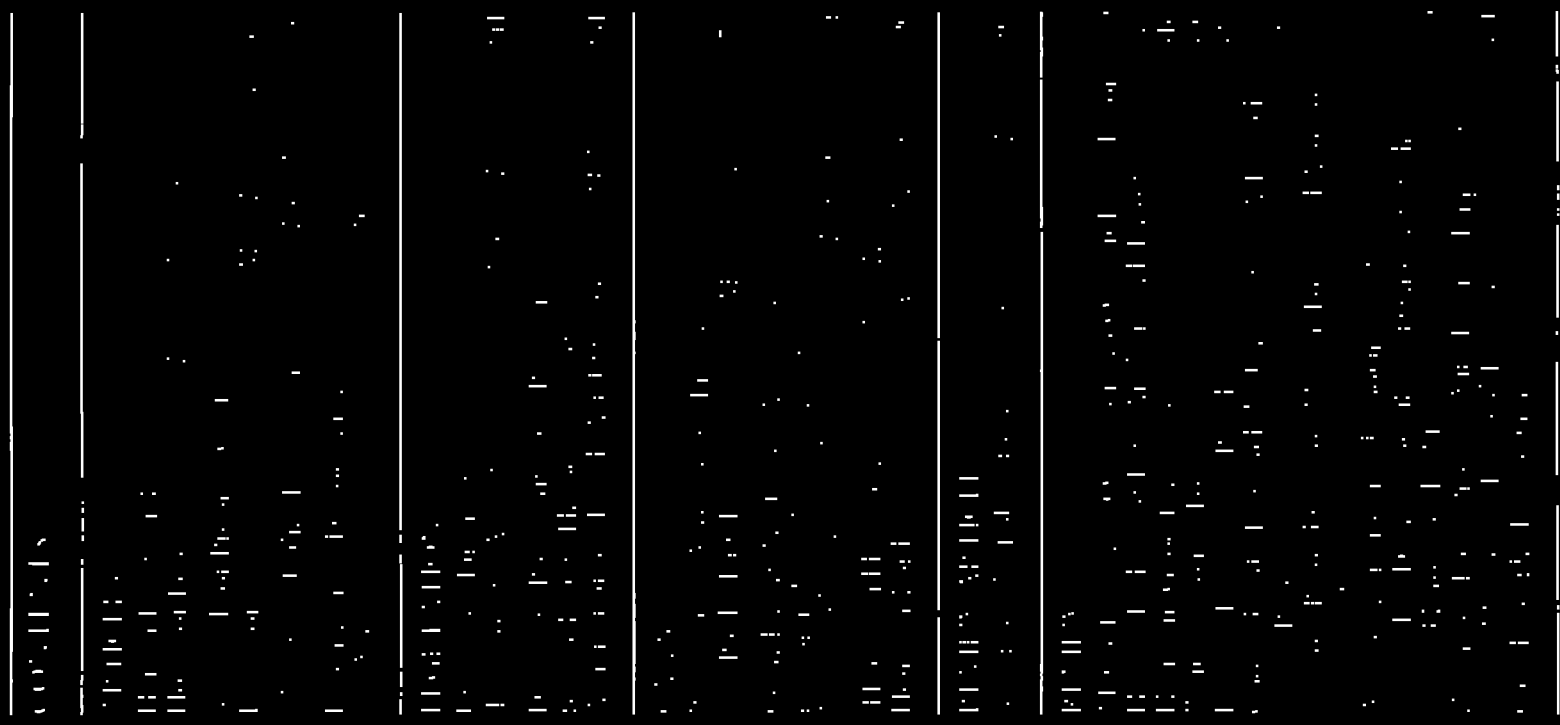
US Mailing

Periodicals postage paid at Rahway, New Jersey.
Postmaster: send address corrections to *The Modern Law Review*, c/o Mercury Airfreight International Ltd Inc., 365 Blair Road, Avenel, NJ 07001, USA (US mailing agent).

Microform

The journal is available on microfilm (16 mm or 35 mm) or 105 mm microfiche from the Serials Acquisitions Department, University Microfilms Inc, 300 North Zeeb Road, Ann Arbor, MI 48106, USA.

0026-7961(199909)62 2 1-Y



VOL. 62, NO. 3 MAY 1999

THE MODERN LAW REVIEW

PUBLISHED FOR THE MODERN LAW REVIEW LIMITED
BY BLACKWELL PUBLISHERS, 350 MAIN STREET, MALDEN, MA 02148, USA

THE MODERN LAW REVIEW

Volume 62

EDITORIAL BOARD

J.S. Anderson
M.R. Chesterman
Lord Dahrendorf of Clare Market
Gerald Dworkin
J.M. Evans
Anthony Giddens
Cyril Glasser
J.A.G. Griffith
Judge F. Mancini
Harvey McGregor
T.M. Partington
I.G. Patel
Lord Wedderburn of Charlton, *Editor Emeritus*
Michael Zander

EDITORIAL COMMITTEE

Julia Black, *Production Editor*
Roger Brownsword, *Case Note Editor*
Hugh Collins, *Articles Editor*
W.R. Cornish
Neil Duxbury, *Reviews Editor*
Judith Freedman
Carol Harlow
Nicola Lacey
Jenny Levin
Martin Loughlin
W.T. Murphy, *General Editor*
R.W. Rawlings
S.A. Roberts
Colin Scott, *Secretary*
R.C. Simpson, *Legislation and Reports Editor*
F.G. Snyder
Erika Szyszczak

Published for The Modern Law Review Limited by Blackwell Publishers.

© The Modern Law Review Limited 1999

ISSN 0026-7961

The Modern Law Review Ltd is incorporated in England as a company limited by guarantee under the Companies Act (Reg No 325 282) and is registered as a charity (Reg No 264072)

The Modern Law Review is published six times a year, in January, March, May, July, September and November by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main St, Malden, MA 02148, USA.

Information for contributors, subscription information and details about microfilm and advertising appear on the inside back cover.

The Modern Law Review website is at *http://www.lse.ac.uk/depts/law/mlr.htm*

Printed and bound in Great Britain by The Alden Press, Oxford
This journal is printed on acid-free paper.

COPYRIGHT: All rights reserved. Apart from fair dealing for the purposes of research or private study, or criticism or review, as permitted under the UK Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing of the Publisher, or in accordance with the terms of photocopying licences issued by organisations authorised by the Publisher to administer reprographic rights, such as the Copyright Licensing Agency and the Copyright Clearance Center

*** FREE COURSEPACK COPYING IN UK TEACHING INSTITUTIONS DURING 1999***

For a trial period only (1 January to 31 December 1999) CLA copyright fees and permissions for the inclusion of articles, notes and reviews from any volume of Modern Law Review in study packs for use by students in UK teaching institutions have been waived. Such copies may be made free of charge during this period provided the institution does not profit from such activites.

CONTENTS

ARTICLES

Foundations of Charity Law in the New Welfare State 333
Michael Chesterman

Rights Brought Home for Children 350
Jane Fortin

Nationalism and the Pathology of Legal Systems: Considering the *Quebec Secession Reference* and its Lessons for the United Kingdom 371
Mark D. Walters

LEGISLATION

The Crime and Disorder Act 1998: Child and Community 'Safety' 397
Christine Piper

'Racist Crime': Racially Aggravated Offences in the Crime and Disorder Act 1998 Part II 409
Maleiha Malik

CASES

Negligent Misstatement, Threats and the Scope of the *Hedley Byrne* Principle 425
Richard Mullender

Pregnancy Discrimination: Equality, Protection or Reconciliation? 435
Michael Wynn

Doctrinal Rationality after *Woollin* 448
William Wilson

REVIEW ARTICLE

The Rhetoric of Precedent and Comparative Legal Research 464
Maurice Adams

REVIEWS

Zucker and Zucker (eds): Medical Futility and the Evaluation of Life-Sustaining Interventions 468
Tonry and Hatlestad (eds): Sentencing Reform in Overcrowded Times 470
Maresceau (ed): Enlarging the European Union – Relations between the EU and Central and Eastern Europe 472
Corder and Maluwa (eds): Administrative Justice in Southern Africa 474
Taggart (ed): The Province of Administrative Law 476
Allen: The Law of Evidence in Victorian England 483
Edwards and Waelde (eds): Law & the Internet 485
Stapledon: Institutional Shareholders and Corporate Governance 486

THE
MODERN LAW REVIEW

Volume 62 May 1999 No. 3

Foundations of Charity Law in the New Welfare State

*Michael Chesterman**

A mission statement for charity law

This is something of a 'back to basics' paper about the peculiar blend of trusts law, tax law and administrative law which we call the law of charities. In general terms, my case is that even though the 'new welfare state' may have provided a very different operating environment for charities, one in which both their significance and their resources have seemingly increased, some basic problems remain in charity law. In fact, the importance of solving these problems satisfactorily is another thing that has increased on account of the arrival of new approaches to welfare provision.

About twenty years ago, in the concluding section of my book, *Charities, Trusts and Social Welfare*, I compiled what I fondly hoped was a cogent left-wing critique of the general structure of English charity law.[1] This critique probably now looks distinctly dated. Lodged within it, but not directly expressed, was a set of loosely-formed ideas about what a satisfactory legal regime for charities should seek to achieve. In outline, the set of ambitions, or 'mission statement', for charity law[2] which I implicitly had in mind was as follows:

- to provide one or more legal structures (trust, company etc) whereby the purposes envisaged by the altruistic originators of publicly beneficial, non-profit-distributing enterprises can be formally laid down;
- to ensure that any processes in which these enterprises, or persons or bodies on their behalf, engage with a view to obtaining funds from government bodies and/or the private sector are carried out in accordance with appropriate standards of honesty and efficiency;

* Faculty of Law, University of New South Wales, Sydney, Australia.
This paper was written for the Colloquium on Foundations of Charity, jointly organised by King's College London and the National Council for Voluntary Organisations in September 1998, and funded by the Modern Law Review and the Economic and Social Research Council. Expert research assistance provided by Angela Seow Wee Tan is gratefully acknowledged I also owe thanks to the Law Foundation of New South Wales for funding this research and to my colleague Philip Burgess for providing critical commentary on a draft.

1 Michael Chesterman, *Charities, Trusts and Social Welfare* (London: Weidenfeld & Nicolson, 1979) chs 14–17.

2 The 'basic characteristics' suggested for charities (not charity law) by the Charity Commission in Annex A to its Consultation Paper, *Framework for the Review of the Register* (London: HMSO, 1998) have some similarities to this 'mission statement' of mine. See too the Report of the Commission on the Future of the Voluntary Sector, *Meeting the Challenge of Change: Voluntary Action into the 21st Century* (London: HMSO, 1996) Part 3.

© The Modern Law Review Limited 1999 (MLR 62 3, May). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

- to ensure that the persons in charge of enterprises of this nature apply their funds and other assets towards fulfilment of their stated purposes, meeting appropriate standards of honesty and efficiency and with some awareness of the desirability of co-operating with other agencies in the same field of welfare provision;
- to provide state support for such enterprises through fiscal privileges, so long as they pursue purposes which can legitimately be expected to benefit a significant group of people who are in need of welfare assistance, and thereby bring about some redistribution of society's resources from the rich to the poor;
- to assure to such enterprises a significant measure of independence from state welfare policy, in order that (a) public-spirited individuals will continue to be motivated to establish and maintain them and (b) any potential which they may have for creative and innovative contributions to welfare provision may be beneficially exploited.

A commentary on my final chapter, published in Australia in 1991,[3] contained the suggestion (with which I agree) that the broad 'conundrum' underlying my critique was that of reconciling the 'private ordering' which takes place through the creation and perpetuation of charitable enterprises with a number of significant 'public interest' considerations. These considerations include efficiency in expenditure of welfare funds, recognition of a hierarchy of needs and co-ordination of welfare activity. The commentary suggested (and again I agree) that total, harmonious reconciliation between these private and public imperatives is impossible, so some degree of compromise is unavoidable.

In this paper, written two decades later on the topic of 'charity law in the new welfare state', what I would like to do is to discuss the extent (if any) to which changes in English charity law and administration over the last twenty years have pursued the fourth and fifth of these broad aims. These aims can be summarised as 'promoting wealth redistribution through tax reliefs' and 'ensuring relative autonomy' respectively.

I recognise that to specify these two aims – particularly the first of them – as appropriate for charity law may be a touch controversial. But it is my belief that in the view of the changes in government welfare policy that are summed up in the phrase 'new welfare state', it is even more important now than twenty years ago that charity law should try to achieve them. Virtually every review of wealth distribution in this age of economic rationalism identifies a widening gap between rich and poor. If the institutions that we call charities are not consistently engaged in seeking to narrow this gap, it is difficult to think of a reason why they should deserve to be called 'charities'.

As my contact with charity law and administration during the last twenty years has almost entirely been in an Australian context, I will unashamedly draw on Australian material when it seems useful.

I will start with a few brief comments on what the arrival of the 'new welfare state' seems to have meant for charities.

3 Terry Carney and Peter Hanks, 'Taxation Treatment for Charities: Distributional Consequences for the Welfare State' in Richard Krever and Gretchen Kewley (eds), *Charities and Philanthropic Organisations: Reforming the Tax Subsidy and Regulatory Regimes* (Melbourne: Comparative Public Policy Research Unit, Monash University and Australian Tax Research Foundation, 1991) 49, 69–71

 © The Modern Law Review Limited 1999

Developments in welfare provision during the last two decades

There is no point in my trying to describe in detail what is 'new' about the 'new welfare state'. I am not qualified to do so, and quite a few others who are supremely qualified to do so have in fact done so. It is enough for me just to repeat a few familiar generalisations which are relevant to the particular points that I want to make about English charity law.

Twenty years ago the voluntary welfare sector, when matched against the welfare state, was consistently viewed as the 'junior partner in the welfare firm'[4] in terms of both overall size and scale of service delivery. It was seen as fulfilling a few distinctive roles appropriate to a 'junior partner'. In outline, these were the roles of advocate of change, pioneer of new services, provider of complementary, additional or alternative services and in some instances sole provider of services.[5]

Since then, as is well known, both the balance between state and voluntary provision of welfare and the nature of the voluntary sector's contribution have changed dramatically. In consequence of the 'storm tides of the Thatcher years',[6] the proportion of the voluntary sector's provision of welfare services has risen markedly. The most noticeable feature of this shift, according to all the commentators, has been the arrival of contracts to deliver welfare services. To a much greater extent than before, the state has engaged voluntary agencies under contract to provide services in circumstances where central or local government agencies or other governmental or quasi-governmental bodies would previously have provided them. It has also outsourced service provision to private commercial enterprises, with which voluntary agencies have had to compete to secure contracts. The voluntary sector has thus had to come to terms with a new 'contract culture'.[7]

This has meant, among other things, that the aggregate of direct cash payments from government to charities and the proportion of such payments to total charity income have increased.[8] So too has the proportion of income which charities receive as fees and charges, whether paid by beneficiaries or by government agencies on their behalf. These developments are of course not uniform within the voluntary welfare sector. The proportion of fees and charges within the income of housing associations is much higher, for instance, than in that of most other species of voluntary organisation.

In addition, the range of fiscal privileges associated with charitable enterprises has increased during the last twenty years, notably for the benefit of donors to charity. There is however reason to believe that *inter vivos* donations by individuals to charity are on the decline, with the National Lottery sometimes cited

4 This is the title of Chapter 19 of David Owen, *English Philanthropy 1660–1960* (Cambridge, Mass: Belknap, 1965), in which the relationship of state welfare and private charity following World War I is described.

5 See eg Report of the Wolfenden Committee, *The Future of Voluntary Organisations* (London: Croom Helm, 1978) 43.

6 Ray Jones, 'Swimming Together. The Tidal Change for Statutory Agencies and the Voluntary Sector' in Chris Hanvey and Terry Philpot (eds), *Sweet Charity: The Role and Workings of Voluntary Organisations* (London: Routledge, 1996) 39, 39.

7 For discussion of some of the legal implications, see Jean Warburton and Debra Morris, 'Charities and the Contract Culture' [1991] Conv 419.

8 See eg Martin Knapp, Eileen Robertson and Corinne Thomason, 'Public Money, Voluntary Action: Whose Welfare?' in Helmut K. Anheier and Wolfgang Seibel (eds), *The Third Sector: Comparative Studies of Nonprofit Organizations* (Berlin: Walter de Gruyter, 1990) 183, 186–199.

© The Modern Law Review Limited 1999

as a contributory factor. The amounts derived from corporate giving and from legacies have increased or more or less stayed the same.[9]

Broad implications for charity law

In the result, the 'junior partner' metaphor for voluntary welfare is still appropriate in the 'new welfare state', though the disproportion between the two 'partners' is a good deal smaller. But for my purposes, the significant developments are (a) the increase in state funding for charities through the mechanism of service delivery contracts and (b) the moves to more favourable fiscal treatment of charities, particularly in relation to donors.

To my mind, the actual impact of these developments has been in many respects detrimental. I realise that it may seem eccentric, if not lunatic, to argue that measures which make charities richer are detrimental. But I have reasons. They stem from the interaction of these measures with two familiar 'chestnuts' of English charity law, which (like the poor) seem destined to remain with us.

The first of these 'chestnuts' is the principle that, to quote a familiar phrase of Lord Cross in *Dingle* v *Turner*,[10] validity as a charitable trust and significant 'fiscal immunity' continue to 'march hand in hand' (the '*Pemsel* principle'). The second is the principle that an enterprise whose purposes include seeking to change some aspect of the law or government policy must be deemed 'political' and therefore non-charitable (the '*Anti-Vivisection* principle').

My argument, in essence, is as follows: so long as these entrenched principles of charity law remain more or less unaltered, their interaction with the phenomena of increased state funding of charities through service contracts and more favourable tax treatment, particularly for donors, impedes, not assists, fulfilment of the two aims for charity law on which I am focusing. This interaction does not promote wealth redistribution through charitable activity and it seriously impairs charity autonomy. It follows, in my view, that reform of these principles – which many people, including myself, have advocated over a long period – is particularly desirable at the present day.

In relation to each of these two principles of charity law – the '*Pemsel* principle' and the '*Anti-Vivisection* principle' – I will now try to explain the steps in my line of argument and to indicate what sort of eminently justifiable changes should be given serious consideration.

Uniform tax benefits for charities and charitable donors (even when they benefit the rich): the '*Pemsel* principle'

The regressive effects of the uniform tax benefits currently available

The charity and the voluntary welfare sectors are not the same. One of the major arguments in my 1979 critique of charity law[11] – and I was not alone in making it at that time[12] – was that they should be brought closer together, at least in the fiscal

9 See eg Graham Moffat, *Trusts Law: Text and Materials* (London: Butterworths, 2nd ed, 1994) 627; Stuart Etherington, 'To the Millennium: The Changing Pattern of Voluntary Organisations' in Hanvey and Philpot, n 6 above, 12–15.

10 [1972] AC 601, 625

11 n 1 above, 349–350, 403–404.

12 See eg Ben Whitaker, 'Minority Report' in National Council of Voluntary Organisations, *Charity Law and Voluntary Organisations* (the Goodman Report) (London: NCVO, 1976) paras 6–10.

 © The Modern Law Review Limited 1999

domain. I argued that the tax reliefs and other 'privileges' associated with charitable status should not be enjoyed, in the name of charity, by organisations, such as 'public' schools and private hospitals, which on any popular understanding of the relevant terms were neither charitable in orientation nor engaged in welfare provision for persons of lesser or middle-range means. They should instead be reserved for organisations whose welfare programmes achieved some measure of wealth redistribution by benefiting the poor and could therefore be genuinely deemed altruistic.[13] Put in these terms, this argument directly attacked the tax law dimension of the House of Lords' decision in *Pemsel*.[14]

As I have indicated, the extensions of tax relief that now particularly concern me affect *inter vivos* donations to charities. Since 1990, one-off gifts of not less than £250 to a charity have attracted relief for income tax on the part of the donor, whether individual[15] or corporate.[16] Since 1991, no upper limit has applied to the amount donated.[17] This form of relief, whether earned through making such a gift or under a voluntary covenant extending over four or more years, exempts the taxpayer from both basic and higher rates of tax.[18] In addition, statutory payroll deduction schemes facilitate the claiming of tax relief by employees.[19]

The overall value of tax reliefs to charities is considerable. A Treasury estimate in 1997 quoted a figure of £1.75 billion.[20]

One clear effect of these extensions of tax benefits for donors to charities is to encourage gifts of significant amounts – ie of £250 or more, if made on a one-off basis – and gifts, whether or not of significant amounts, made by people in the higher tax brackets, who now obtain relief at both basic and higher rates. To the extent that these changes have in fact enhanced the scale of donations, they have allowed to be diverted to charities yet more private money that might otherwise have been brought into public revenue as income tax. And the important underlying point, even more important in the light of these changes, is that the state has no means of controlling the amount of tax revenue foregone on gifts or covenants to a charity or indeed income or capital profits realised by a charity. If the prescribed conditions are satisfied, the relevant tax relief can be claimed as of right. What the charity receives is 'blanket, automatic, no-strings attached, open-ended aid'.[21] It follows also, of course, that there can be no control over the total amount of tax foregone in favour of the charity sector as a whole, though this is not the issue of specific interest here.

In addition, these extensions of donors' tax relief favour richer donors, and the charities which they choose to support, in two distinct ways. First, the new relief in respect of one-off gifts only operates if the amount given is at least £250. Only relatively wealthy people make one-off gifts of this magnitude. Secondly, the relief granted now extends to the income tax that would otherwise be paid by the donor

13 For an argument in similar terms (though lacking the specific concern with tax policy), see Alan Ware, 'Meeting Needs through Voluntary Action: Does Market Society Corrode Altruism?' in Alan Ware and Robert E Goodin (eds), *Needs and Welfare* (London: Sage, 1990) 185, 197–202.

14 *Commissioners of Income Tax* v *Pemsel* [1891] AC 531.

15 Finance Act 1990, ss 25, 26, as amended by Finance (No 2) Act 1992, s 26 and Finance Act 1993, s 67.

16 Income and Corporation Taxes Act 1988, ss 338, 339; Finance Act 1990, s 26.

17 Finance Act 1991, s 71.

18 Income and Corporation Taxes Act 1988, ss 347A–350, 660, 683; Finance Act 1989, s 59.

19 Income and Corporation Taxes Act 1988 s 202 and Finance Act 1993, s 68

20 HM Treasury, News Release, 'Charity Taxation Reviewed', 2 July 1997.

21 Stanley S. Surrey, 'Federal Income Tax Reform: The Varied Approaches Necessary to Replace Tax Expenditures with Direct Governmental Assistance' (1970) 84 HLR 352, 385; and see discussion in Rick Krever, 'Tax Deductions for Charitable Donations. A Tax Expenditure Analysis' in Krever and Kewley, n 3 above, 1, 18–19.

© The Modern Law Review Limited 1999

at higher rates. While in many cases the charities which rich donors favour may genuinely assist the poor, the potential of this 'upside-down' factor to direct further state financial support towards fee-charging elitist institutions which by virtue of charging high fees operate almost entirely for the benefit of rich people – for example, elite independent schools[22] or expensive private hospitals – is quite obvious.[23] By contrast, limits can be and of necessity are placed on the amounts paid to individual social welfare charities by or on behalf of the state under granting programmes or as fees for providing welfare services.

In this instance, the balance between private 'ordering' and 'public interest' that I spoke of earlier is clearly tipped towards private 'ordering'. The decision of a wealthy private donor to make a substantial gift to a charity which benefits only people in the middle and higher income ranges automatically drags with it a significant sum of public money, irrespective of the public interest considerations in issue.

These arguments highlight the weakness of one of the significant propositions appearing in the case-law to justify charitable status for enterprises such as independent schools or private hospitals. I am referring to the statement of Lord Wilberforce in *Re Resch's Will Trusts*[24] to the effect that in establishing the requisite public benefit (which may be 'indirect' as well as direct) to qualify for charitable status, a private fee-charging facility in a field such as health care or education may argue that it relieves pressure on equivalent state facilities.

This principle was invoked in 1982 in *Joseph Rowntree Memorial Trust Housing Association* v *Attorney-General*.[25] It was held by Peter Gibson J that the provision by a housing association of residential long leases for elderly people, irrespective of their means, was within the fourth class of charitable purposes because the accommodation was specifically designed to meet the particular needs of 'the aged'. He confirmed prior authority, including the decision in *Resch*, to the effect that the phrase 'aged, impotent and poor people' in the Elizabethan Preamble should be read disjunctively. His judgment includes a statement, similar to that of Lord Wilberforce, that the provision of such leases conferred 'a clear public benefit' on the community in so far as the tenants could 'defer the time when they may fall on to state services'.[26]

In conformity with this approach, the Charity Commission has on occasions approved objects clauses which envisage a lessening of the burden on the National Health Service through the provision of 'mainstream' private health facilities.[27] Yet this principle is contradicted by the Commissioners in their published statements relating to poverty relief charities. They have said in several of their Annual Reports that the funds of these charities should not be used simply to relieve the statutory responsibilities of state welfare authorities.[28]

22 In a study published in 1989, it was estimated that the value of tax concessions to independent schools was probably between five and ten per cent of their fee income: Graham Moffat, 'Independent Schools, Charity and Government' in Alan Ware (ed), *Charities and Government* (Manchester: Manchester UP, 1989) 190, 200–202. This study, of course, predated the tax changes of the early 1990s.

23 Krever, n 21 above, 19–21; see too Paul R. McDaniel, 'Federal Matching Grants for Charitable Contributions: A Substitute for the Income Tax Deduction' (1972) 27 Tax L Rev 377, 390–392.

24 [1969] 1 AC 514, 544 (PC). See critical discussion in n 22 above, 199; Jeffrey Hackney, 'The Politics of the Chancery' [1981] CLP 113,121–122.

25 [1983] Ch 159. See critical discussion in Richard Nobles, 'Selling Charity' (1983) 46 MLR 782.

26 n 25 above, 167. The second phrase in quotation marks had appeared in a letter from the Charity Commissioners to the housing association's solicitors, and was cited with approval by Gibson J.

27 Moffat, n 9 above, 664–665.

28 Annual Reports of the Charity Commissioners (London: HMSO) 1967, paras 17–20; 1975, para 29, 1991, Appendix A(b)

 © The Modern Law Review Limited 1999

One cannot always assume, however, that private provision in an area such as health will in fact relieve state expenditure. In some instances, private providers make significant use of state facilities, without paying full value for them.[29] Also, any 'public benefit' arising because the relevant charitable organisation reduces pressure on equivalent state services may be substantially offset by the financial support accruing to it through tax concessions. As I have just emphasised, the scale of this support cannot be controlled to any degree by the Treasury. Furthermore, the organisation in question may in addition be receiving cash grants.[30] Accordingly, in the absence of empirical justification, the use of this argument, based on 'relief of the burden on state welfare', in order to support charitable status for private facilities in realms such as health, housing and education is highly questionable. As I have just indicated, it is considered dubious by the Charity Commission in the context of relief of poverty.

In any event, even if the private facility does in fact confer a net benefit on the state, the overall level of 'handout' to each of the beneficiaries coming from the state (through fiscal exemptions and cash grants) and from the funds raised otherwise by the charity may well be distinctly greater than occurs under any equivalent state scheme, even after the value of the fees and charges paid by the beneficiaries is taken into account. A further ruling in the *Rowntree* case shows how charity law's interpretation of 'public benefit' leaves scope for this. Gibson J held that various schemes for providing long leases which the association had proposed for approval by the association were charitable in law and valid. This was the case even though the tenants, on exercising a right to surrender their leases, might derive a profit because the lease had increased in value during their tenancy. Because any such profit was 'at most incidental to the charitable purpose',[31] it did not matter that the schemes provided, in essence, an opportunity for investment by the beneficiaries as well as accommodation specially designed to meet their specific needs.

There is in fact no clear limit to the level of generosity, indeed luxury, to which the benefits dispensed by a fee-charging charitable organisation may aspire. In discussing charity law, one should never forget the swimming-pool that Marlborough College did not need. (My reference here is to *Re Geere (No 2)*,[32] in which a bequest of £10,000, a substantial sum in 1954, to furnish the College with 'a proper swimming bath and hot water baths' was held charitable and valid, but had to be applied *cy-près* because the College already possessed these facilities.)

By contrast, less prosperous charitable organisations which provide welfare services, without charge, to less well-off people will inevitably be confined to dispersing meagre benefits. A striking illustration appears in the Charity Commissioners' guidelines indicating how charities for the relief of poverty may engage in 'topping up' state social security benefits.[33] The measures listed include 'the provision of items [either outright *or, if expensive and appropriate, on loan*, (emphasis added) such as] (a) furniture, bedding, clothing, food fuel, heating appliances; (b) washing machines *for widows with large families* or radio or television sets *for the lonely, bedridden or housebound*' (emphasis added). (There

29 Moffat, n 9 above, 676.

30 For discussion in relation to independent schools, see eg n 22 above, 202.

31 n 25 above, 176.

32 [1954] CLY 388; *The Times*, 19 May 1954, Chancery Division, Harman J; for brief commentary, see n 1 above, 335; Hubert Picarda, *The Law and Practice Relating to Charitable Trusts* (London Butterworths, 2nd ed, 1995) 43.

33 Annual Report of the Charity Commissioners for 1991 (London: HMSO) Appendix A(b).

© The Modern Law Review Limited 1999

are shades here of the notoriously moralistic comment of Romer LJ in *Re Cole*[34] that 'I cannot regard the provision of television sets, etc, for the benefit of such persons as juvenile delinquents and refractory children in [a local authority home for such people] as coming within any conception of charity which is to be found in the preamble'.)

In addition, such charities are increasingly operating under relatively strictly-drawn contracts to deliver services on behalf of the state, where again the scale of individual benefits will be distinctly meagre. This is, as I have already said, the 'new role' for voluntary welfare.

The problem remains, therefore, that the automatic harnessing, ever since *Pemsel*, of tax privileges to the preamble-based trusts law definition of 'charitable' contradicts my charity law aim of 'fiscally promoting wealth redistribution'. I hope that I have managed to demonstrate that this contradiction has been accentuated by the tax changes of the last two decades.

One familiar solution for this problem, frequently mentioned,[35] is to exclude from the legal definition of charitable those bodies whose activities cannot in practice confer any significant benefit on people of limited means. Another, which I wish to explore now, is to apply differential tax treatment to different classes of charities. Predictably, my contention is that those charities which genuinely benefit poor people should receive the most favourable tax treatment. This should, I think, extend particularly to the donors to such charities. The enlarged reliefs for donors which England introduced in the early 1990s should clearly not, in my view, have operated indiscriminately amongst all charities.

A model for tax reform

In Australia, a number of tax laws, including particularly the provisions governing income tax relief for donors, have in fact broken away from the principle that the pursuit of 'charitable' purposes, as defined by the majority judges in *Pemsel*, should be the sole password to fiscal privileges so far as philanthropic or welfare-oriented bodies are concerned. There is instead a two-tier set of exemptions, in which 'eleemosynary' charities, and the donors to such charities, receive the most favourable treatment.

This process of breakaway began as a consequence of a case in 1926 called *Chesterman* v *Federal Commissioner of Taxation*.[36] Here the Privy Council reversed a decision of the High Court of Australia[37] to the effect that 'charitable', in the context of an exemption from estate duty, should be construed rather along the lines of the minority judgments in *Pemsel* – that is, according to the word's popular meaning, covering only purposes involving relief of poverty. The Privy Council insisted that the definition prescribed by the majority in *Pemsel*, embracing the traditional four categories derived from the Elizabethan preamble, should apply instead.

The response of the Australian Parliament in 1928 was to amend the estate duty legislation so as to grant exemptions to bequests to a different category of recipients, that of 'public benevolent institution'. This concept is now employed in

34 [1958] Ch 877, 888 For criticism, see n 1 above, 335–337, 355, Picarda, n 32 above, 44.

35 See eg n 12 above, paras 6–10, n 22 above, 217.

36 [1926] AC 128 (PC). The case arose out of a claim for exemption from estate duty for the assets in a testamentary trust (the Peter Mitchell Trust), under which prizes for physical, moral and literary excellence were to be awarded to young men irrespective of their means The first-named trustee was my grandfather.

37 (1923) 32 CLR 362

© The Modern Law Review Limited 1999

lieu of 'charitable' as a criterion for relief from a number of Commonwealth and State taxes in Australia, including payroll tax, fringe benefits tax, sales tax and income tax payable by donors.[38] The term 'charitable' continues to play a fiscal role also: for example, the investment and trading income of charitable institutions and of funds established by will or instrument of trust for 'public charitable purposes' is exempt from income tax.[39] Under some provisions,[40] exemptions may be claimed on either ground. In the preparation of a governing instrument of any trust, association or other body operating in a philanthropic or welfare field in Australia, it is now commonplace to try to frame objects which will establish it both as a charitable body and as a 'public benevolent institution'.

To qualify as a 'public benevolent institution', an organisation must be engaged in relief of 'poverty, suffering, distress or misfortune'[41] or of 'destitution or helplessness',[42] or of those 'who are in need of [relief] and are unable to care for themselves', namely 'the poor, the sick, the aged and the young'.[43] In early cases, it was suggested that if fees were to be charged, they would have to be purely nominal or be levied on a sliding scale according to the means of the beneficiary.[44] Later authorities have however held that an institution for the relief of sickness or some other distress or misfortune may be 'benevolent' even if it charges fees at a level approaching the value of services rendered.[45] A body which is established, funded and controlled by government and which therefore operates in essence as an arm of the state cannot qualify,[46] but a body which is funded but not controlled by government agencies may do.[47]

In recent years, the phrase has been held to cover an institute established to assist the mentally ill[48] and a co-operative society providing housing, education and employment opportunities for needy Aboriginal people within a defined area and also promoting the development of their culture and identity.[49] On the other hand, the Royal Society for the Prevention of Cruelty to Animals (RSPCA) has been held outside the category,[50] even though the court recognised that its activities were beneficial to mankind more or less for the same reasons as appear in the cases holding animal welfare bodies to be charitable.[51]

A significant problem for my overall argument is the fact that income tax relief for donors in Australia, while not available, as I have said, for donors to all charitable bodies, is not restricted to gifts to public benevolent institutions. The relevant statutory provision, section 30-45 of the Income Tax Assessment Act 1997 (Cth), also makes it available for gifts to several other limited categories of

38 This last and most significant exemption is conferred by the Income Tax Assessment Act 1997 (Cth), s 30-45.

39 Income Tax Assessment Act 1997 (Cth), s 50-5.

40 For example, Local Government Act 1919 (NSW), s 132(1)(d), dealing with liability for rates.

41 This phrase comes from the judgment of Dixon J in the first High Court case to consider the question of interpretation, *Perpetual Trustee Co Ltd* v *Commissioner of Taxation (Cth)* (1931) 45 CLR 224, 233–234.

42 *ibid* 232, *per* Starke J.

43 *ibid* 235–236, *per* McTiernan J.

44 *ibid* 235, *per* Evatt J; *Lemm* v *Federal Commissioner of Taxation* (1942) 66 CLR 399.

45 *Commissioner of Pay-roll Tax (Vic)* v *Cairnmillar Institute* (1992) 92 ATC 4307.

46 *Metropolitan Fire Brigades Board* v *Commissioner of Taxation* (1990) 27 FCR 279.

47 *Tangentyere Council Inc* v *Commissioner of Taxes (NT)* (1990) 90 ATC 4352 (reversed on other grounds (1992) 107 FLR 470).

48 n 45 above.

49 *Maclean Shire Council* v *Nungera Co-operative Society Ltd* (1995) 86 LGERA 430 (Court of Appeal, NSW).

50 *Re Royal Society for the Prevention of Cruelty to Animals, Queensland Inc* [1993] 1 Qd R 571.

51 *ibid* 583, *per* Thomas J.

© The Modern Law Review Limited 1999

institution which have charitable purposes – for example, public educational institutions – or to any one of a large number of individual listed charities (which include, for instance, marriage guidance councils and the RSPCA).

Yet the important concept that charitable status should not be the sole password to tax privileges for bodies of a philanthropic nature and their supporters remains within Australian tax law. The equally significant category of 'public benevolent institution' more clearly reflects the popular meaning of 'charitable' (and the minority views in *Pemsel*)[52] in its focus on relief of the needs of poor or disadvantaged people. It provides a model for any common law jurisdiction which wishes, wholly or partly, to confine tax reliefs to bodies within this narrower category and/or to individuals or corporations making gifts or bequests to such bodies.

I must also concede, however, that it is not a perfect model in terms of my 'mission statement' for charity law. It is open to criticism[53] on the ground that it implements a distinctly conservative view of charitable activity, according to which only traditional 'relief' organisations can qualify as 'public benevolent institutions'. This means that organisations whose activities are directed at 'changing the circumstances which either create or aggravate poverty or distress'[54] are excluded. So too, it would seem, are organisations which engage in research, publication and advocacy with a view to assisting and representing other bodies which dispense relief. Pursuant to the first of these principles, the status of 'public benevolent institution' was denied in 1985 to the Australian Council of Social Service, Australia's peak welfare council (the equivalent, more or less, of the National Council for Voluntary Organisations).[55] In the view of one judge, the latter principle also provided a ground for this ruling.[56]

A further ground of criticism is that the court's initial instinct to deny the status of 'public benevolent institution' to bodies which charge more than nominal fees has given way to a recognition that fee-charging is not incompatible with 'benevolence'. For reasons that will have become apparent, I believe this to have been an unfortunate development in the case-law.

If Australia were to set about redefining 'public benevolent institution' within in its current tax laws, or if England were to adopt the same general approach to framing tax reliefs, some changes to the definition should therefore, I believe, be made. I have no quarrel with the use of phrases such as 'poverty, suffering, distress or misfortune' and 'destitution or helplessness' to define the circumstances in which potential beneficiaries must find themselves. But nineteenth century concepts of 'relief' should be jettisoned in favour of a more liberal requirement that the purposes of a public benevolent institution must simply entail some form of benefit to a section of the public[57] which qualifies as beneficiaries. The current

52 ie those of Lord Halsbury LC and Lord Bramwell, n 14 above, 542, 564–566. On the similarity between the concept of 'public benevolent institution' and the popular notion of charity, see eg *Ashfield Municipal Council* v *Joyce* [1978] AC 122, 137 (PC), n 46 above, 283.

53 See eg n 3 above, 68–69, 75–76. For criticism from a rather different perspective, see E. Blake Bromley, 'Problems in Defining Charity' in Australian Association of Philanthropy Inc, *Australian Charity at the Crossroads* (Melbourne: Australian Association of Philanthropy, 1994) 38.

54 *Australian Council of Social Service Inc* v *Commissioner of Pay-roll Tax* (1985) 1 NSWLR 567, 571, *per* Priestley JA; and see *Marriage Guidance Council of Victoria* v *Commissioner of Pay-roll Tax* (1990) 90 ATC 4770

55 *Australian Council of Social Service*, n 54 above.

56 *ibid* 568–569, *per* Street CJ.

57 This means, it will be noted, that benefit solely to 'poor relations' or 'poor employees' would not suffice, irrespective of decisions such as *Dingle* v *Turner* (n 10 above) to the effect that trusts for these classes of beneficiaries may fall within the legal definition of 'charitable purpose'.

© The Modern Law Review Limited 1999

requirement, for instance, that a body cannot be a 'public benevolent institution' unless it directly provides 'relief' to the relevant class of beneficiaries is unnecessarily restrictive. In addition, the earlier rulings which apparently insisted that no significant fees could be charged should be reinstated.

The question whether a body might legitimately receive especially favourable tax treatment as a public benevolent institution if its activities are directed to 'changing the circumstances which either create or aggravate poverty or distress' brings me directly to the second issue that I have singled out for discussion. This is the status of 'political' purposes within the law of charity.

'Political' is not 'charitable' (even in a democracy): the '*Anti-Vivisection* principle'

The background to my discussion of this well-known principle of charity law is the emergence of the 'contract culture' associated with the 'new welfare state'. Essentially my argument is that the autonomy and the innovative potential of charities are now threatened from two sides.[58]

On one side, their engagement in contracts with state authorities to deliver welfare services robs them of their freedom to decide for themselves how they might most effectively apply their resources for the benefit of their chosen constituency. They are effectively, or at least potentially, bound to acquiesce in and implement the welfare policies of the governmental or quasi-governmental body with which they are in contract. It has been suggested that the general tendency of the service contract model is indeed to favour larger, longer-established, more specialised and generally more cautious charities.[59]

On the other side, the limits which charity law imposes on 'political activities', particularly since the 1982 case of *McGovern* v *Att Gen*,[60] significantly inhibit those charities which are not too tightly bound by service contracts from publicly challenging those policies. The legal divide between political and charitable purposes has indeed become deeper as a result of Slade J's ruling in the *McGovern* case that Amnesty International was not a charity.

So far as charities operating in the United Kingdom are concerned, the important aspects of Slade J's judgment are twofold. First, it reinforced the rule that advocating any change in the law was a political and therefore non-charitable purpose. Secondly, and most significantly for my purposes, it confirmed the extension of this rule to pressing for changes in government policies or administrative decisions.[61] The legal justifications provided were familiar: that courts have no means of judging whether changes in laws, government policies or administrative decisions would be for the public benefit, and that to

58 For a forceful argument along these lines relating specifically to unemployment relief charities, see Maria Brenton, *The Voluntary Sector in British Social Services* (London: Longman, 1985) 93–100.

59 See eg Nicholas Deakin, 'Government and the Voluntary Sector in the 1990s' (1991) 12 (3) *Policy Studies* 11, Marilyn Taylor, 'The Changing Role of the Nonprofit Sector in Britain: Moving Towards the Market' in Benjamin Gidron, Ralph M. Kramer and Lester M. Salamon (ed), *Government and the Third Sector* (San Francisco: Jossey-Brass, 1992) 147; Marilyn Taylor, 'The Canary in the Coalmine. Issues Facing the Voluntary Sector' (1996) 8 *Social Policy Review* 40.

60 [1982] Ch 321.

61 It is generally assumed that opposing changes in the law is also 'political' and therefore non-charitable: see *Re Hopkinson* [1949] 1 All ER 346, 350, Picarda, n 32 above, 161–162.

© The Modern Law Review Limited 1999

decide that the law is not right as it stands would usurp the functions of the legislature.[62] The judgment has been frequently criticised.[63]

The Charity Commissioners' Guidelines on the matter[64] somewhat alleviate the situation for campaigning charities by indicating how they may make the most of the principle that political activity which is merely ancillary to one or more charitable purposes does not endanger charitable status. But many charities dealing with poverty and disadvantage remain convinced that they cannot operate effectively without a significant degree of campaigning. They are inhibited in doing this because it is very difficult to tell in advance whether a campaign will be deemed by the Commission to have gone beyond merely ancillary or subordinate conduct.[65]

The recent valuable review of the attitudes of 24 legal systems to this issue, conducted by Anita Randon and Perri 6,[66] shows clearly that the English rules, along with those of most other 'charity law countries', are firmly situated at the restrictive end of the spectrum. Those countries which do not adhere to a common law concept of 'charity' appear remarkably relaxed about voluntary welfare organisations engaging in campaigns about welfare issues.

The authors of this article also set out an argument which has been noticeable in critical commentary on the subject during the last couple of decades[67] and which I would now like to develop further. Its starting-point is the proposition that the participation of individual citizens and non-governmental groups in public agitation to change laws and government policies is an integral and beneficial element of democracy. This notion underpins legal guarantees of freedom of speech, whether they take the form of express statements in a country's constitution (as in the USA), or arise as constitutional implications (as in Australia) or are mandated by instruments of international law (such as Article 10 of the European Convention on Human Rights). A number of recent cases in England, notably the decision of the House of Lords in *Derbyshire County Council* v *Times Newspapers Ltd*,[68] have acknowledged the importance of freedom of public discussion of political matters within a democratic society.

When, however, in a recent English case an opportunity arose to take account of these considerations in a context analogous to charity law, it was not taken up. In the 1996 case of *R* v *Radio Authority, ex p Bull*,[69] the British Section of Amnesty International (AIBS) challenged a decision of the Radio Authority banning

62 n 60 above, 333–337, 339–340, citing *Bowman* v *Secular Society* [1917] AC 406, 442, *per* Lord Parker and *National Anti-Vivisection Society Ltd* v *Inland Revenue Commissioners* [1948] AC 31, 62–63, *per* Lord Simonds. The extension to advocacy of changes in government policy was applied in *Webb* v *O'Doherty, The Times*, 1 February 1991, Chancery Division, Hoffmann J (injunction to restrain students' union, a charity, from spending funds to support a campaign to stop the Gulf War).

63 See eg Richard Nobles, 'Politics, Public Benefit and Charity' (1982) 45 MLR 704; Friedl Weiss, 'Quot Homines Tot Sententiae or Universal Human Rights: A Propos *McGovern* v *The Attorney-General*' (1983) 46 MLR 385, 385–392; D. Lee, 'Charities and Politics – The Perennial Problem' (1984) VI Liverpool L Rev 49; Caroline J Forder, 'Too Political to be Charitable?' [1984] Conv 263; also sources cited in nn 67, 72 below.

64 Charity Commission for England and Wales, Leaflet CC9, *Political Activities and Campaigning by Charities* (London: HMSO, 1997).

65 For a recent affirmation of this point, see Francine Bates and Jill Pitkeathley, 'Standing Up to be Counted: Campaigning and Voluntary Agencies' in Hanvey and Philpot, n 6 above, 82

66 'Constraining Campaigning: The Legal Treatment of Non-profit Policy Advocacy Across 24 Countries' (1994) 5(1) *Voluntas* 27.

67 See *ibid* 38; and eloquent passages in C.E F. Rickett, 'Charity and Politics' (1982) 10 NZULR 169, 170–173; Charles Rickett, 'Politics and Cy-Près' [1998] NZLJ 55, 62

68 [1993] AC 534, where it was held that a local council, being a governmental body, could not sue for defamation at all.

69 [1996] QB 169; [1997] 2 All ER 561 (CA)

© The Modern Law Review Limited 1999

licensed radio stations from broadcasting its advertisements. The ban was imposed on the ground that the advertisements fell within a prohibition in section 92(2)(a)(i) of the Broadcasting Act 1990 covering advertisements 'inserted by or on behalf of any body whose objects are wholly or mainly of a political nature'. The objects of AIBS were clearly political in the broader sense of seeking to change laws and/or government policies, but were not party-political. The same could be said of the content of a proposed advertisement, relating to human rights abuse in Rwanda and Burundi, which was before the court. In upholding the ban, both the Divisional Court and the Court of Appeal relied on the broad definition of 'political' which Slade J had applied in *McGovern* v *Att Gen*[70] in ruling that the Amnesty International Trust was not a charity.

The specific circumstances of this case provided a good opportunity for injecting into a judicial definition of 'political' some acknowledgment of the importance of freedom of public discussion of political matters within a democratic society. AIBS argued that a legislative policy against party-political advertising was understandable at least, if not positively beneficial (though interestingly in *Australian Capital Television Pty Ltd* v *Commonwealth*,[71] the Australian High Court had held in 1992 that a legislative ban on all forms of broadcast political advertising generally, including party-political advertisements, was contrary to an implicit principle of freedom of political discussion within the Australian Constitution). But it maintained that the broader definition of 'political', embracing advocacy of any changes in law or policy, even when these would promote respect for human rights, should not have been transplanted unquestioningly from charity law into the law regulating media communications. This contention was, however, rejected.

As a valuable commentary on the case in *Public Law* points out,[72] the decision in *R* v *Bull, ex p Radio Authority* was unfortunate from the point of view of both of these branches of law. In relation to charity law, the authors state:[73]

> Those cases which established that charitable status should not extend to purposes which are political belong to an earlier social era. The fundamental objection to according such status, namely that the law is incapable of judging whether a change in the law or government policy is good or bad, is clearly spurious. In a relativistic age and a mature democracy the law should be able to uphold as charitable objects which are diametrically opposed to each other, provided that they are for the 'public benefit' in the view of a sizeable body of adherents. ... [T]he law of charities ... should be allowed to evolve so as to encompass the observance of human rights, whether in this country or overseas, and the honouring of international obligations which have been entered. The very fact that the U.K. has acceded to such obligations should be conclusive evidence of benefit to the community.

A comparable approach to softening the rigour of the *McGovern* decision was in fact adopted in a judgment in 1997 in the Supreme Court of New South Wales. The case, *Public Trustee* v *Att Gen (NSW)*,[74] concerned the objects of a non-governmental organisation (now defunct) called the Federal Council for the Advancement of Aborigines and Torres Strait Islanders. Citing earlier High Court authority,[75] Santow J drew a distinction[76] between trusts for purposes which are

70 n 60 above.
71 (1992) 177 CLR 106. The Divisional Court in *Bull* (n 69 above, 181, *per* Kennedy LJ) cited a passage from the judgment of Mason CJ (at 145) in this case.
72 J. Stevens and D.J. Feldman, 'Broadcasting Advertisements by Bodies with Political Objects, Judicial Review, and the Influence of Charities Law' [1997] PL 615.
73 *ibid* 622.
74 (1997) 42 NSWLR 600.
75 *Royal North Shore Hospital of Sydney* v *Att Gen (NSW)* (1938) 60 CLR 396, 426, *per* Dixon J.
76 n 74 above, 607–608.

© The Modern Law Review Limited 1999

'contrary to the established policy of the law', which are political and automatically fail, and trusts whose object is to 'introduce new law consistent with the way the law is tending', which 'may survive in Australia as charitable'. He suggested, though he did not have to decide, that a trust with 'a modest agenda for legislative change' might fall into the latter category where 'the change is to promote anti-discrimination legislation conforming to the trend in legislative policy and reinforced by norms of international law ...'.[77]

The important question in charity law is of course not whether advocacy of change in laws or policies should merely be permissible but whether it is of benefit to the public, indeed sufficiently beneficial to attract fiscal privileges. In terms of their strict legal effect, guarantees of freedom of speech – which, it should be said, is also one of the 'norms of international law',[78] as indeed of European law[79] – generally address only the former issue. Their chief effect is to proclaim, subject to numerous exceptions and qualifications, that speech should not attract legal sanctions. But in at least one legal context in which the impact of notions of free speech has been much discussed in recent years, it has in fact been held that public discussion of matters of government or politics or other matters relating to the 'public life' of the community is positively beneficial to the community. The relevant area of law is civil defamation.

In England, the significant decision is as recent as July 1998. In *Reynolds* v *Times Newspapers Ltd*,[80] a defamation case instituted by a former Prime Minister of a neighbouring country (Ireland), the Court of Appeal held that the defence of common law qualified privilege could protect the makers of defamatory communications to the public at large in a much wider range of situations than had previously been recognised. The long-established justification for this defence is that it serves 'the common convenience and welfare of society'.[81] It operates to protect defamatory communications, even when they are untrue, if they are made without malice by a person having a legal, moral or social duty or interest to publish the material and the recipients have a reciprocal duty or interest to receive the material.

For present purposes, the important passage in the Court of Appeal's judgment in *Reynolds* is as follows:[82]

> We do not for an instant doubt that the common convenience and welfare of a modern plural democracy such as ours are best served by an ample flow of information to the public concerning, and by vigorous public discussion of, matters of public interest to the community. By that we mean matters relating to the public life of the community and those who take part in it, including within the expression 'public life' activities such as the conduct of government and political life, elections and public administration, but we use the expression more widely than that, to embrace matters such as (for instance) the governance of public bodies and companies which give rise to a public interest in disclosure, but excluding matters which are personal and private, such that there is no public interest in their disclosure.

In reformulating and extending the defence of qualified privilege to reflect these considerations, the Court of Appeal stated that the 'nature, status and source' of the published material and 'the circumstances of the publication' should be taken into

77 *ibid* 617: see also at 619.
78 For example, under Article 19 of the International Covenant on Civil and Political Rights.
79 Under Article 10 of the European Convention on Human Rights.
80 [1998] 3 WLR 862.
81 *Stuart* v *Bell* [1891] 2 QB 341, 346 (CA), *per* Lindley LJ, drawing upon *Toogood* v *Spyring* (1834) 1 CM & R 181, 193, *per* Parke B
82 n 80 above, 909.

 © The Modern Law Review Limited 1999

account. In the words of the judgment, this constituted 'the circumstantial test'.[83] False defamatory material taken from 'a government press release, or the report of a company chairman, or the speech of a university vice-chancellor' and published to the world at large would be more likely to attract the defence (other things being equal) than 'the statement of a political opponent, or a business competitor or a disgruntled employee'.[84] By the same token, the Court said, 'it is one thing to publish a statement which has been so far as possible checked, and quite another to publish it without such verification as was possible and as the significance of the statement called for'.[85]

An alternative formulation had been developed almost exactly one year earlier by the High Court of Australia. This was in the case of *Lange* v *Australian Broadcasting Corporation*,[86] another defamation case brought by the former Prime Minister of a neighbouring country (New Zealand). The High Court held that the criterion for determining whether statements relating to 'government and political matters' should be protected by the qualified privilege was whether there had been 'reasonableness of conduct' in making the publication. The Court of Appeal in *Reynolds* suggested that its 'circumstantial test' might not be very different from this Australian test of reasonableness.[87]

These judgments clearly imply that courts are capable of distinguishing between two categories of statement which contribute to the 'flow of information to the public concerning, and [to] vigorous public discussion of, matters of public interest to the community'.[88] These are respectively statements that pass the 'circumstantial test' (or, in Australia, are published 'reasonably') – which means that publication of them to the world at large sufficiently serves 'the common convenience and welfare of society' to attract the defence of qualified privilege – and statements that fail this test (or, in Australia, are not published 'reasonably').

Returning, after this brief excursus, to the topic of charities, it could not be doubted that public statements urging changes in law or government policy, which the *McGovern* decision[89] treats as 'political', are within the broad concept of 'matters of public interest to the community', as conceived by the Court of Appeal in *Reynolds*. With this in mind, consider now the legal implications of a hypothetical report, accompanied by a press release, issued by a well-established and well-respected unemployment relief charity.

The report and press release vigorously attack the responsible Minister for promulgating regulations lowering the level of youth unemployment benefits, allege that she has systematically misinterpreted data provided to her about the cost of maintaining benefits at pre-existing levels and urge the Government to adopt a much more sympathetic welfare policy towards unemployed young people. The report is based on reasonably careful investigation of the background to the new regulations, and both the charity employees who prepared it and the journalists who disseminate its arguments honestly believe its contents to be true.

Suppose that the Minister takes defamation proceedings against the charity and against the media organisations which published these allegations. In this situation,

83 *ibid* 899.
84 *ibid* 909–910
85 *ibid* 910.
86 (1997) 189 CLR 520. In New Zealand, see *Lange* v *Atkinson* [1998] 3 NZLR 424 (Court of Appeal), where qualified privilege was held to protect any defamatory publication, if honestly made, on a political matter without any further 'circumstantial test' or criterion of reasonableness. The Court of Appeal in *Reynolds* rejected this approach (n 80 above, 907).
87 *ibid* 910.
88 *ibid* 909.
89 n 60 above.

© The Modern Law Review Limited 1999

the defendants might well be able to show that the 'circumstantial test' set out in *Reynolds* was satisfied and that the defence of qualified privilege was therefore available to them. If the court agreed, publication of the report and press release would be implicitly held to serve 'the common convenience and welfare of society'.

On the other hand, opponents of the stance on unemployment adopted by the charity could use these documents in support of an argument to the Charity Commission that the charity trustees were pursuing political, not charitable, aims. In terms of the Guidelines issued by the Commissioners, they would appear in particular to breach prohibitions against expressing opposition to particular political parties.[90] The Commission could be invited to take steps against the trustees to recover any expenses associated with the report and press release, or indeed to consider deregistration of the charity.[91] As already indicated, the legal principles underpinning this argument include the proposition that advocacy of a change in the law or in government policy cannot constitute a charitable purpose because a court has no means of judging whether the change would be for the public benefit, or because the court cannot usurp the functions of the legislature (the law cannot 'stultify itself'[92]) by holding that any change would be beneficial.

In strict terms, this evident contradiction can be reconciled, though only just. The principle of defamation law states that society derives benefit from an 'ample flow of information to the public concerning, and ... vigorous public discussion of, matters of public interest' such as a past or proposed change in law or government policy. On the other hand, the principle of charity law is not, in its terms, concerned with information flow or public discussion, but with whether the change is or is not beneficial.

Yet in practical terms this distinction is illusory. In the charity cases, the courts regularly and misleadingly describe the impugned trusts as trusts 'to change the law', as if nothing more than a decision by the trustees would immediately enact the relevant legislation. At most, however, they are no more than trusts to communicate to relevant government authorities, and to the electorate, arguments in favour of changing the law – ie to convey information to government and to the public concerning, and engage in public discussion of, this issue of public interest.

It is wholly contradictory for the courts to say that in certain circumstances, in the context of defamation law, this activity serves 'the common convenience and welfare of society', but that it cannot in any circumstances be adjudged to be for the public benefit, in the context of charity law. In the former situation, courts must now be prepared to discriminate, through use of the 'circumstantial test', between those communications to the public which promote 'the common convenience and welfare of society' and those which do not. Yet in the latter, they are constrained by authority to declare themselves incapable of drawing any such distinction. The justifications offered are distinctly archaic. They include, for instance, the notion that it is for the legislature alone to say whether a change in the law would be beneficial. This looks very like Lord Reid's 'fairy tale'[93] that judges never make the law, but only declare it. They also betray scant regard for the role of popular opinion in a democracy.

A further anomaly is that, within the limits of their powers, the Commissioners do make such a distinction in their Guidelines. In fact, some of the criteria applied

90 See n 64 above, paras 19, 46, 47, 50.

91 *ibid* paras 71–75. As is well known, Oxfam has on occasions considered it necessary to tone down its political activities in response to warnings by the Commission: see eg n 33 above, 19–20.

92 *National Anti-Vivisection Society*, n 62 above, 50, *per* Lord Wright.

93 See 'The Judge as Lawmaker' (1972) 12 JSPTL 22, 22.

© The Modern Law Review Limited 1999

by them in determining whether a charity's advocacy of changes in law or government policy crosses the line from legitimate to illegitimate political activity resemble those of the 'circumstantial test' outlined in the *Reynolds* case. The Guidelines state, for instance, that a charity may carry out such advocacy within the range of its charitable purposes by presenting a 'reasoned' written memorandum to the Government, or by seeking to influence public opinion 'by well-founded reasoned argument'.[94]

The overriding legal principle which the Guidelines seek to elaborate is however that political activity by a charity can only ever be ancillary to its charitable, non-political purposes. Adhering to this principle is, as I have said, difficult for charities because, as the Commission concedes, 'the dividing line between proper debate in the public arena and improper political activity is a difficult one to judge'.[95] The argument which I am putting here would go a long way towards reducing those difficulties. It is that participation in public debate about changes in law or policy relating to an area of charitable activity should be recognised as beneficial within a democratic society and, within limits, charitable in its own right. Judicial acceptance of this argument would significantly loosen the constraints on campaigning charities.

This means, of course, that organisations with significantly political orientations could claim tax privileges. This notion is not total heresy: under present law,[96] gifts to political parties attract exemption from inheritance tax and consequential relief under the capital gains tax regime, to the same extent as gifts to charities. But the arguments which I have put forward in the preceding section for differentiating between charities in framing tax reliefs for the charity sector would of course apply to political charities. If under a differential tax regime, charities which engaged genuinely in activities assisting the poor and disadvantaged obtained more favourable treatment than other charities, they should not lose this fiscal advantage merely because one of their purposes was to campaign for changes in law or government policy within their field of operation.

94 n 64 above, para 36.
95 *ibid* para 10.
96 Inheritance Tax Act 1984, s 24, Taxation of Chargeable Gains Act 1992, s 260.

© The Modern Law Review Limited 1999

Rights Brought Home for Children

Jane Fortin*

The background

The incorporation of the European Convention for the Protection of Human Rights and Fundamental Freedoms into domestic law was described, albeit unsympathetically, as a measure which would have 'a seismic impact on the people of this country. It is part of the bulldozing of the constitutional landscape of the United Kingdom'.[1] The object of this article is to consider the extent to which children will benefit from this change. The government's plan to make the European Convention on Human Rights part of our domestic law, marks a radical change of approach to traditional thinking on how best to protect the rights and freedoms of British citizens. The view was formerly that our innate freedoms were properly protected by a democratically elected parliament and that they existed unless and until they were expressly abrogated by common law or legislation. Furthermore, it was considered that to adopt a written constitution, as in France and the USA, would endanger underlying liberties, since only those listed would gain recognition.[2] Now the government considers that it is not enough merely to assert the theoretical existence of citizens' fundamental liberties. Instead, their constitutional rights require formal guarantees and direct modes of protection.[3]

The United Kingdom's erstwhile refusal to incorporate into domestic law any list of rights has meant that although successive governments have ratified many international human rights treaties, these have had no direct effect on English law. Nevertheless, the list of rights embodied by the European Convention has already become a relatively familiar one. Indeed, despite its not being part of English law, it has had considerable influence on the development of law here. This was partly because the courts have been entitled to assume, when interpreting domestic legislation, that the legislature had not intended it to be inconsistent with any of the United Kingdom's obligations under the Convention.[4] Of greater significance however, was the existence of a powerful mechanism for the Convention's interpretation and enforcement. Since 1959, when the European Court of Human Rights was established, individuals have had a means of obtaining redress for domestic

* School of Law, King's College, London.

I am most grateful to my colleague Conor Gearty for reading and commenting on earlier drafts of this article.

1 Gerald Howarth MP, HC Deb vol 306 col 838 16 Feb 1998

2 The United Kingdom's approach to protecting rights and the arguments for and against introduction of a Bill of Rights are discussed by D. Feldman, *Civil Liberties and Human Rights in England and Wales* (Oxford: Clarendon Press, 1993) 60–88 See also C. Gearty 'The United Kingdom' in C. Gearty (ed), *European Civil Liberties and the European Convention on Human Rights: A Comparative Study* (Dordrecht: Martinus Nijhoff, 1997). Gearty, at 65–83, provides a comprehensive discussion of the legal and political background to the campaign to incorporate the Convention. See also 84–103 for a discussion of the decisions reached by the European Court of Human Rights involving the United Kingdom.

3 *Rights Brought Home: The Human Rights Bill* Cm 3782 (London: HMSO, 1997) esp paras 1.16–1.19.

4 See eg the efforts of Lord Oliver in *Re KD (a minor) (ward: termination of access)* [1988] 1 All ER 577, 587–589 to reconcile English law with the terms of the European Convention. Since the United Kingdom is bound by international law to observe all treaties it has ratified, the courts must presume that all domestic legislation is intended to comply with the terms of all such treaties

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

infringements by taking their complaints to Strasbourg. This special feature has undoubtedly enhanced the Convention's reputation for being one of the world's foremost human rights instruments.[5] In the past, the United Kingdom has often been required to answer charges that its laws have infringed individuals' rights protected by the Convention and to amend them, when found to be in breach of its provisions.

Once the Human Rights Act 1998 is fully implemented, matters will change quite radically. Arrangements will be made enabling individuals to invoke their Convention rights in the domestic courts in this country, rather than having to apply to Strasbourg. Put in such mundane terms the change does not appear particularly dramatic. Nevertheless, the provision which will, in truth, have a 'seismic impact' on the law here is the provision making it unlawful for a 'public authority'[6] to act in a way which is incompatible with Convention rights. This means that all arms of government will be forced to consider the human rights of all those affected by their day to day operations. Summarised very briefly the new system will enable an individual to apply to the domestic courts, arguing that the government or any public authority is infringing his or her rights under the Convention and that he or she is a victim of that unlawful act.[7] If the court considers the claim to be legitimate, it may grant 'such relief or remedy, or make such order, within its powers as it considers just and appropriate'.[8] In some cases, the infringement will be attributable to the terms of primary legislation, which itself takes no account of the Convention right or rights claimed. In such a case, the courts[9] will be empowered to declare that legislative provision incompatible with the terms of the Convention.[10] It will then be for the government and parliament to put matters right if they choose, through a 'fast-track' procedure established to amend the law as swiftly as possible, to bring it into conformity with the Convention.[11] More crucially, the common law will also be open to challenge, since like any other 'public authority' the courts must also refrain from acting 'in a way which is incompatible with a Convention right'.[12] Consequently if litigants can satisfy the courts that the existing principles of the common law infringe their Convention rights, the courts must adjust those principles through their decisions, in a way that ensures compatibility. The implications of such a change are enormous and will be discussed in more concrete terms below.

5 Until recently the decision making process was a complicated one. Petitions were investigated first by the European Commission of Human Rights and in the event of its being satisfied that the petitioner had exhausted all domestic remedies and that the petition had some merit, it attempted to effect a reconciliation between the opposing parties. Should this fail, either the Commission or the State Court would then refer the case to the European Court of Human Rights to obtain a legal ruling. When Protocol 11 came into force on 1 November 1998, the system was simplified and the part-time Commission and Court were replaced by a single full-time Court which decides both on the admissibility and merits of each case. See Articles 28 and 29 of the European Convention on Human Rights, as amended by Protocol 11.

6 Defined by the Human Rights Act (HRA) 1998 s 6 (3) as including '(a) a court or tribunal, and (b) any person certain of whose functions are functions of a public nature . .'.

7 *Ibid* s 7(1).

8 *Ibid* s 8(1); s 8(2) and (3) makes clear that the remedy might include the award of damages.

9 The power to make such declarations is limited to the House of Lords, Court of Appeal and High Court: *ibid* s 4(5).

10 ie by making a 'declaration of incompatibility': *ibid* s 4(2).

11 'Remedial action' will then be taken (see s 10 and Sched 2) by the government minister taking responsibility for that particular piece of legislation, designed to ensure that it is amended speedily to remove the incompatibility in question. Amending legislation should be put in place within 120 days – see Sched 2 para 4 (4).

12 See s 6(1) and (3). This point is made and discussed by C. Gearty, in 'The Impact of the New Human Rights Act', in A. Boyle and A. Roux (eds), *Human Rights and Scots Law* (Oxford: Hart, 1999 forthcoming).

© The Modern Law Review Limited 1999

The European Convention on Human Rights and children's claims

Now everyone, regardless of age, is to have specific entitlement to all those rights listed in the European Convention on Human Rights. Children will inevitably benefit from the heightened rights consciousness accompanying incorporation. But for them, this change could achieve far more, if it is utilised with perspicacity, taking account of its innate limitations. In practical terms, those advising children may now have acquired the means of proving Onora O'Neill wrong. This writer once asserted that since an appeal to children's rights has little chance of empowering them, their remedy is to grow up.[13] In her view, if children are too young they will be wholly unable to respond to the appeal, if they are old enough to respond, they will probably be well on the way to majority and to ending the disability and dependence peculiar to childhood. Incorporation of the European Convention into domestic law may refute this depressing prognosis. Children complaining that their rights under the European Convention are being infringed will no longer have to wait until they are well into adulthood before they receive a decision from Strasbourg.[14] Indeed, the new judicial power to grant declarations of incompatibility[15] may achieve rapid improvements in the existing legislation affecting children. Thus the terms of any existing legislation considered by the courts to be incompatible with children's rights under the European Convention will be capable of swift amendment. New legislation will also be drafted with Convention rights firmly in mind.[16] Government departments with a particular responsibility for children's issues will be required to assess carefully any draft legislation which directly or indirectly affects children in every context, to ensure its compatibility with their Convention rights. As commented above, children will also be able to challenge the principles of the common law, since the courts must not interpret these 'in a way which is incompatible with a Convention right'. If satisfied that these principles offend a child's rights, the courts must adjust them through their decisions, in a way that ensures compatibility. Children's legal advisers and the judiciary will obviously need to rapidly acquire a detailed knowledge of the provisions of the Convention and the forty years of case law associated with its interpretation.

The rights listed in the European Convention reflect the determination of citizens living in post-war Europe to prevent a recurrence of the atrocities and persecution experienced prior to and during World War II. Its provisions broadly protect an individual's personal life and his or her need to live life to the full, and secure the rights enjoyed by him or her when in conflict with the authorities and the law. The Convention emphasises those freedoms without which the concept of individual autonomy would be valueless and which therefore need special protection from state interference. Many of the Articles support the notion that the freedom to determine one's own life style is the key to life in a civilised society. For example, Articles 9, 10 and 11 specifically single out and protect the right to, respectively,

13 O. O'Neill, in P. Alston, S. Parker and J. Seymour (eds), *Children, Rights and the Law* (Oxford: Clarendon Press, 1992) 38–39.

14 It has taken, on average, five years to get an application into the European Court and has cost an average of £30,000.

15 HRA 1998, s 4, discussed above. But see also s 3(1) which makes it clear that such declarations will only be made in the last resort and that the courts should first attempt to interpret existing legislation in a way which is compatible with the Convention.

16 See s 19 which requires any government bill to be accompanied by a written ministerial statement asserting that the provisions of the proposed legislation are compatible with Convention rights.

© The Modern Law Review Limited 1999

freedom of thought, conscience and religion; freedom of expression, including the right to receive and impart information; and the freedom of peaceful assembly and association with others. Applications alleging infringements of these rights may be strengthened by a claim under Article 14, which overarches all the others – that the applicant is being treated in this manner because of his or her membership of a particular sect or group, and that this amounts to discrimination.[17]

Although these fundamental rights are theoretically available to children, there is only a relatively small body of Convention case law illustrating an appreciation that they are as valuable to children as to adults. Indeed the paucity of Convention case law dealing with complaints of this nature brought by or on behalf of children suggests that neither children, nor those advising them, have often thought in terms of the law either giving or depriving them of such freedoms. Consequently, in the past, the European Commission and Court were less likely to be presented with applications brought from children's own perspectives. *Handyside* v *United Kingdom*[18] provides a good example of claims involving children being considered from a particularly adult standpoint. There the European Court had considered the extent to which the right to freedom of expression was breached by the United Kingdom government when, through police action, it seized and destroyed copies of a sex education book for children. This question was assessed from the viewpoint of the adult publishers of the book and not from that of the children themselves.[19] The European Court of Human Rights accepted that the United Kingdom government was fully entitled to argue that *The Little Red Schoolbook* endangered children's moral development and that therefore it was entitled to interfere with the publishers' freedom of expression. Had the application in *Handyside* been brought by an adolescent, the European Court would obviously have been forced to consider a child's own right to freedom of expression, including the right to receive information under Article 10. It was arguable that a book of this kind provided adolescents with essential information about maturing into adult society and that the government ban directly interfered with their rights as individuals. Incorporation of the Convention will provide the opportunity for older British children to claim rights of this kind on their own behalf before the domestic courts. The only qualification is that, like all claimants, they must show an active violation of their rights, as individuals, rather than claiming a violation of the theoretical rights of children, as a group.[20]

In order to exploit the Convention rights constructively for children, their advisers must themselves become fully au fait with Convention case law and the manner in which complaints regarding children have been received in Strasbourg. This is because under section 2 (1) of the Human Rights Act 1998 domestic courts are obliged to take account, where relevant, of the way in which the European Commission and Court have interpreted the Convention in the past. The domestic courts are likely to follow the practice of the European Court, which is normally to follow Convention case law, unless there are cogent reasons for not doing so. The domestic courts are not, however, locked into an anachronistic interpretation of the Convention since the Court itself has stressed that it is free to depart from previous decisions in order 'to ensure that the interpretation of the Convention reflects

17 Article 14: 'The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion ...'.

18 (1976) 1 EHRR 737.

19 See G. Van Beuren, *The International Law on the Rights of the Child* (Dordrecht: Martinus Nijhoff, 1995) 135.

20 ie the applicant must be able to establish that he or she is a 'victim' of the unlawful act. See HRA 1998, s 7(1) and (7)

© The Modern Law Review Limited 1999

societal changes, and remains in line with present-day conditions ...'.[21] Nevertheless, although the domestic courts will doubtless refuse to follow Convention case law if they consider it to be historically inappropriate, it will obviously be enormously influential. The importance of Convention precedent may not, however, always augur well for British children. The European Commission and Court has at times appeared to adopt the common assumption that children's interests are identical with those of their parents. Certainly, as the discussion below will show, it has not demonstrated particularly clearly a full understanding that promoting children's rights may be impossible, without acknowledging that they have an existence independent from their parents.

There are four further aspects of the European Convention which undermine its ability to promote children's rights very effectively. The first is its narrow scope. The European Convention relates only to civil and political rights, which in themselves may be of no value to those too poor or ill to benefit from them. In contrast, the UN Convention on the Rights of the Child not only provides children with fundamental freedoms, such as freedom of expression, but requires states to make resources available to fulfil their social and economic rights. It sets out in some detail the extent to which governments are expected to take positive steps towards promoting these rights, for example by guaranteeing their right to the highest attainable standard of physical and mental health, thereby enabling them to take full advantage of their civil and political rights.

The second limitation is that although the European Convention, by applying to all individuals, regardless of age, does not exclude children from its ambit, it was not designed to cater for them, as a distinct group. The Convention neither refers to the special nature of childhood nor attempts to protect children's rights, as distinct from those of adults. Consequently, unlike the UN Convention on the Rights of the Child, there is no intention to promote the full range of children's needs: for example their need to special protection and care; their need to fulfil their potential through a broad education, and through good health and a reasonable standard of living. Nor is there any specific aim to protect them from their innate vulnerability to abuse and to economic and sexual exploitation.

The third limitation is linked to the second, in so far as the Convention takes no account of the differences between children and adults. Consequently, efforts to ensure that procedures cater for children's special vulnerabilities may be hampered by the need to conform to the relatively inflexible standards of international human rights law. Thus whilst it may be entirely appropriate to develop a non-contentious and constructive approach to dealing with questions relating to children's welfare, the procedural framework underpinning such an approach may fail to match up to the ideals of fairness required by the Convention.[22]

Lastly, and perhaps of greatest concern is the absence of any guidance on how to reconcile parents' rights to freedom from state interference with children's own rights to fulfil their potential and to develop a degree of independence from their parents, as they mature into adulthood. As discussed below, it is this weakness in the European Convention which may leave children without protection against the power of authoritarian parents. Indeed, there is the danger that the Convention will be used to promote parents' rights at the expense of children's rights. [23]

21 See *Cossey* v *United Kingdom* [1991] 2 FLR 492, 501 This point is made and discussed by C. Gearty, 'Human Rights in Practice: some preliminary reflections on the Government's Human Rights Bill' (1998) unpublished paper, at 9.

22 See below in connection with the children's hearing system in Scotland.

23 This concern is discussed by J. Fortin in *Children's Rights and the Developing Law* (London: Butterworths, 1998) 51–56.

© The Modern Law Review Limited 1999

Children's rights under the European Convention

Children and family life

As noted above, many of the Articles represent concerns about the basic freedoms considered essential to civilised life. Nevertheless, at least until late adolescence, because children's lives are centred on their homes and families, adult carers stand interposed between them and the state. Consequently for many children the most important form of protection the Convention can provide relates to their family life with their parents.

Family life and immigration practice

Those seeking to promote children's rights in a family context will normally turn to Article 8 for assistance.[24] It recognises that the concept of freedom requires states to respect an individual's right to physical and psychological privacy, linked with the power to make his or her own family arrangements without interference. Children, like adults, are entitled to respect for their privacy. Since an important aim of Article 8 is to protect family life from state interference, it may provide children with protection from some of the harsher aspects of immigration law and practice. For example, the immigration authorities may sometimes seek to deport an adult who has entered this country and established a home and family here. Immigration officials may not take full account of what will happen to that adult's children if their family is split up by his or her deportation. Even if the children themselves have a right to remain here, the immigration authorities may assume that the simple solution is for them to accompany the deported parent abroad, without adequately considering the extent to which this will disrupt the children's own family rights. Indeed, there is, in some aspects of immigration practice, 'a risk that the child will be viewed only as an appendage of the parents'.[25]

The European Commission has acknowledged that such actions may infringe a child's own right to family life, as protected by Article 8, if his or her parent's deportation will break up their family.[26] But although Article 8 undoubtedly can provide children caught up in immigration disputes with valuable protection, its limitations should not be overlooked. Unlike the provisions of the UN Convention on the Rights of the Child, Article 8 is not designed with children particularly in mind. It contains no specific direction to consider the welfare of any minor members of the family, far less to make the child's interests paramount. Furthermore, Article 8 is not absolute and the parent's removal may be justified under Article 8(2), if deemed to be proportionate with the legitimate aim pursued by the government in taking such action.[27] In other words, a child's rights, like those of any adult, must be balanced against the government's right to maintain a firm system of immigration control.[28]

24 Article 8(1) provides that 'Everyone has the right to respect for his private and family life, his home and his correspondence'.

25 *R* v *Secretary of State for Home Department, ex p Gangadeen, R* v *Secretary of State for Home Department, ex p Khan* [1998] 1 FLR 762, per Swinton Thomas LJ, at 784.

26 See *Uppal* v *United Kingdom (No 1)* (1981) 3 EHRR 391 and *Uppal* v *United Kingdom (No 2)* (1981) 3 EHRR 399. Three generations of the family claimed that their right to respect for family life had been infringed by the planned deportation, the grandparents, the parents and the children. See also *Poku* v *United Kingdom* (1996) 22 EHRR CD 94.

27 Article 8(2) provides 'There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety ... or for the protection of the rights and freedoms of others'.

28 See *Abdulaziz* v *United Kingdom* (1985) 7 EHRR 471.

© The Modern Law Review Limited 1999

It is clear that the immigration authorities are fully aware of the government's obligations to heed the aims of Article 8 when operating its immigration policy in relation to families.[29] But to date, if their practice was challenged for violating its terms, they were entitled to argue, as the domestic courts themselves had done, that the Convention was not part of domestic law and that, in any case, states are entitled to operate within a wide margin of appreciation.[30] Furthermore, if deportees and, or, their children complained to Strasbourg about their treatment, it depended entirely on the circumstances of each individual case whether, by expecting a child to accompany the parent being deported, or by deporting a parent and leaving the child in this country, the Home Secretary would be deemed by the European Court to be in breach of Article 8.[31] Incorporation may encourage the adoption of a more consistent policy to take greater account of children's own rights in immigration cases. This aside, the immigration officers and tribunals will all fall within the definition of 'public authority'[32] and must therefore act in accordance with the terms of the Convention. Furthermore, a planned removal of a parent from this country deemed by the domestic courts to be in breach of the rights of parent and or child under Article 8, will, in itself, justify a finding of breach of the Human Rights Act and therefore of illegality.

A degree of concern has also been voiced over the absence of an appropriate and properly organised system for assisting unaccompanied minors seeking refugee status. This should be available from the time they arrive and for ensuring that they have an appropriate form of representation before the Immigration Appeal Tribunals.[33] Children objecting to these aspects of immigration practice might seek to persuade the domestic courts that they infringe Article 6.[34] Sometimes immigration officers dealing with such children do not believe their assertions that they are minors, so that a few end up detained by the immigration service in young offender institutions or even prisons.[35] The government has recently indicated that this should only occur in the most exceptional circumstances.[36] Certainly the experiences that these children undergo whilst detained strongly indicate that such practice is inhumane and therefore in breach of their rights under Article 3.[37] Additionally, unless the detention is extremely brief, it might be

29 See Deportation Policy Guidance (DP/3/4/5/96) Home Office. Article 3 might also be relevant, but only if the family is returned to an extremely harsh regime. *Chahal* v *United Kingdom* (1996) 23 EHRR 413.

30 eg *R* v *Secretary of State for the Home Department, ex p Lucy Ouma* [1997] Imm AR 606. See also *R* v *Secretary of State for Home Department, ex p Gangadeen, R* v *Secretary of State for Home Department, ex p Khan* (1998) 1 FLR 762.

31 eg *Berrehab* v *The Netherlands* (1989) 11 EHRR 322; *Abdulaziz* v *United Kingdom* (1985) 7 EHRR 471; *Poku* v *United Kingdom* (1996) 22 EHRR CD 94.

32 See HRA 1998, s 6 (3).

33 See W. Ayotte, *Supporting Unaccompanied Children in the Asylum Process* (London: Save the Children, 1998) at paras 2.3.2, 3.8 1 and 3.8.2. See also A. Jones, 'United Kingdom Immigration Policy and Welfare Implications: Children and Young People' (1998) Unpublished seminar paper for IPPR, at 30–32. In *Bektas Orman* v *Secretary of State for the Home Department* [1998] Imm AR 224, 227 Judge Pearl suggested that a different approach to dealing with an unaccompanied child seeking refugee status should be considered in order to ensure that he or she is always accompanied by a Refugee Council Panel Adviser at all stages of the immigration process, including all aspects of the appeal.

34 Article 6 secures an entitlement to a fair and public hearing by an independent tribunal, in the determination of civil rights. Since a deportation decision may impinge on a child's right to family life, which, unlike a right to remain in a country rather than face deportation, is a civil right under the Convention, the procedural guarantees of Article 6 should be available. See n 108 below.

35 A. Jones n 33 above, 5–27. See also W. Ayotte n 33 above, para 3.91

36 *Faster, Firmer, Fairer – A Modern Approach to Immigration and Asylum*, Cm 4018 (London: HMSO, 1998) para 12.6.

37 W. Ayotte n 33 above, paras 3.9.4, 4.3 16 and 4.4.4.

© The Modern Law Review Limited 1999

challenged for infringing their rights under Article 5(4) which provides those deprived of their liberty with the right to a speedy hearing to test the legality of their detention.[38] At first sight such detention could also be challenged for infringing the detailed provisions in the UN Convention on the Rights of the Child protecting refugee children.[39] Regrettably on ratification of that Convention the United Kingdom entered a reservation entitling it to apply whatever legislation to such children 'as it may deem necessary from time to time'.

Children's family life and parents' privacy

Convention case law reflects a very conservative approach to family life and to the parental role. As discussed below, the European Commission and Court have seemed often to assume that parents can be trusted to control and discipline their children as they think fit. This leads to certain difficulties. First it fails to deal adequately with conflicts between parents and children. Second it down-grades the protection against state interference that children should be entitled to expect from the Convention. Third, there is the implicit assumption that parents' claims under the Convention will always benefit their children, when in fact this may not be the case.

The first difficulty is exacerbated by the fact that none of the Articles were designed to deal with the complex interplay between parents' powers and children's rights. Indeed, even taking account of the European Court's recent decision on corporal punishment,[40] it can be argued that none protects children very adequately from repressive parents. Article 8 appears to assume that as long as families are adequately protected from undue state interference, the regulation of family life can be entrusted to its adult members. Equally, the Convention case law has failed to tackle the problem of how to balance a child's own right to privacy against his or her parents' right to run their household as they please. Indeed, the Commission rejected the complaints of a 14 year-old Dutch runaway that her rights under Article 8 had been infringed by the actions taken by the welfare authorities to return her home. Even if they had infringed her rights under Article 8, this was justified by their need to protect her health and morals under Article 8(2).[41] It will be for the advisers of British children to persuade the domestic courts that these early authorities were decided at a time when little serious attention had been given to children's capacity to make sensible decisions for themselves.

Article 5 suffers from the same problem so far as children are concerned. Whilst it was clearly intended to protect citizens' liberty from state interference, it was not designed to determine the extent to which parents are entitled to restrict their children's liberty. The European Court acknowledged, in a decision reached in the late 1980s, that in some circumstances, parents might infringe a child's rights to liberty under Article 5, depending on his or her age.[42] Nevertheless, the majority of the Court held, contrary to the view of the European Commission, that the rights of a 12 year-old boy under Article 5 had not been breached by his placement and

38 Recent research indicates that some children have been detained for quite substantial periods, even up to a year, *ibid* paras 3.9.1 and 4.3.16.

39 Article 22.

40 *A* v *United Kingdom* (*human rights: punishment of child*) [1998] 2 FLR 959 discussed in more detail below.

41 *X* v *Netherlands* (1974) (Application No 6753/74) ECHR 1975–76 Vol 1–3 Decisions and Reports 118. But see *X* v *Denmark* (1978) (Application No 6854/74 ECHR) 1977–78 Vol 7–9 Decisions and Reports 81 There the European Commission refused to accede that the Danish welfare authorities had infringed the parents' rights under Article 8 by refusing to force their 14 year-old daughter to return home against her will.

42 *Nielson* v *Denmark* (1989) 11 EHRR 175.

© The Modern Law Review Limited 1999

retention in a psychiatric ward on the request of his mother. Whilst the majority considered that he was still at an age where it was normal for decisions to be made by a parent, even against his wishes, the implication was that at an older age this might not have been reasonable.[43] The Court's assumption that a twelve year old with no mental impairment had no right to object to his mother's wishes seems completely contrary to the spirit of the Convention, with its respect for human dignity, as expressed in its preamble. Van Beuren strongly criticises the Court's failure to consider such situations from the child's point of view or to tackle parents' apparent ability to exploit their position of power over their children.[44] In her view, the majority decision of the European Court left those children who are deprived of their liberty, on the authority of their parents, wholly unprotected by the Convention. It will obviously be for children's advisers to urge that such an interpretation of Article 5 is now completely anachronistic.

This cautious approach to children's rights against their parents leads to the second difficulty referred to above. This is that children's own rights against state interference may be diluted. Thus although Article 8 appears to protect children from interference by an over-authoritarian government, such a government might argue that children's rights thereunder are materially weaker than those of adults, by virtue of their being under their parents' control. For example, on the face of it, children under the age of 10 could use Article 8 to challenge the 'local child curfew schemes' established under sections 14 and 15 of the Crime and Disorder Act 1998, by claiming breach of their right to lead their private lives without state interference. But the government might argue that a child's right to wander the streets at will is already constrained by lawful parental control, as acknowledged by the European Commission.[45] This argument need not necessarily be fatal to attempts to protect children, since authoritarian state interference might be thwarted by parents complaining that their own rights have been infringed under Article 8. Parents could therefore argue that the child curfew schemes would infringe their rights to determine for themselves how and where their children spend their time. Although parents' success with such a claim would obviously benefit their children, it is regrettable that children's rights to family privacy against state interference require protecting in such an indirect manner

Finally, there is the third difficulty – namely that parents' rights under the Convention are sometimes assumed to be identical to their children's rights. Indeed, in this respect, Article 8 can be something of a two-edged sword for children. Whilst it provides a tool for parents to strengthen their own rights to family privacy, it may not always adequately safeguard children's rights to their own private and family lives, free from *parental* interference. For example, British parents have successfully challenged the methods used by the government when removing their children into state care.[46] Their success led to changes in the domestic law which strengthened parents' procedural rights to oppose the actions of social service departments involved in child protection work.[47] Such amendments have strengthened family autonomy and may certainly benefit children if

43 *ibid*195. The three dissenting judges were all convinced that the boy's rights under Article 5 had been infringed.

44 See G. Van Beuren, n 19 above, 213 See also 73–74.

45 *X* v *Netherlands* n 41 above.

46 eg *W, R, O, B and H* v *United Kingdom* (1987) 10 EHRR 29, 74, 82, 87 and 95

47 eg the local authority power to pass parental rights resolutions by way of administrative rather than judicial process (as authorised by Children Act 1948, s2) was abolished and disputes over the extent to which parents can enjoy contact with children who are the subject of care orders must now be determined through formal judicial proceedings (Children Act 1989, s34)

© The Modern Law Review Limited 1999

they prevent their being taken away from their parents unnecessarily or from losing touch with their parents. Nevertheless, domestic courts may now be confronted by parental challenges buttressed by appeals to Article 8. They should remember that as long as the law maintains a respect for the parents' family life by involving them in the decision-making processes, the authorities may well be able to show that the steps taken to protect a child were 'necessary' to protect him or her within paragraph 2 of Article 8.[48]

Fortunately the domestic courts can also take account of the European Court's growing perception that children too have rights which may require balancing against those of their parents.[49] Thus, for example, the European Court has accepted that although taking a child into state care should be regarded as a temporary measure disrupting family life for as short a time as possible, domestic governments cannot *always* be expected to force a child to return to live with his or her birth parents, even against the child's own wishes. Indeed, it has indicated some sympathy with domestic governments' arguments that the child has his or her own rights, as guaranteed by the UN Convention on the Rights of the Child.[50] Furthermore these decisions and other recent case law suggest an increasing acceptance by the Commission and Court that the UN Convention on the Rights of the Child is an influential text regarding the interpretation of children's rights under the European Convention.[51]

Children's family life and the use of corporal punishment

As noted above, neither Article 8 nor indeed any of the other Articles in the Convention cope very adequately with attempts to ensure that children are protected from over-authoritarian parents within the privacy of the home. This weakness has undermined attempts to outlaw the use of corporal punishment in British homes. As is well known, the continued lawfulness of corporal punishment in private homes and independent schools was strongly criticised by the Committee on the Rights of the Child.[52] Challenges employing the Human Rights Act 1998 may eventually achieve its complete abolition in private homes.[53] The use of the European Convention certainly had a dramatic success in gradually curtailing corporal punishment in schools. Rulings by the European Court against the United Kingdom for allowing this form of punishment, thereby infringing the rights of parents[54] and children[55] prompted legislation in 1986 outlawing its use in all state schools.[56] A later decision of the European Court[57] led to further legislation in

48 *W* v *United Kingdom*, n 46, above, para 64.

49 eg *Johansen* v *Norway* (1997) 23 EHRR 33 and *Hokkanen* v *Finland* [1996] 1 FLR 289. See the discussion by J Fortin, n 23 above, 51–56. See also the much earlier decision of the European Commission in *X* v *Denmark* n 41 above.

50 *Hokkanen* v *Finland ibid* 304–305 (paras 57–58).

51 eg *A* v *United Kingdom* n 40 above, para 22. Discussed below

52 Consideration of Reports of States Parties *United Kingdom of Great Britain and Northern Ireland*, CRC/C/15/SR 205 (Geneva. Centre for Human Rights, 1995) at paras 16 and 32. Children's rights in relation to the use of corporal punishment is discussed by J. Fortin, n 23 above, ch 10.

53 See below.

54 *Campbell* v *Cosans* (1982) 4 EHRR 293 where parents successfully argued that the use of corporal punishment in schools infringed their own rights under Article 1 of Protocol 2.

55 *Tyrer* v *United Kingdom* (1978) 2 EHRR 1. See also the decision of the European Commission in *Y* v *United Kingdom* (1994) 17 EHRR 238, paras 44–45. These decisions established that the use of corporal punishment might infringe children's rights under Article 3.

56 Education (No 2) Act 1986, ss 47 and 48. This prohibition also applied to independent schools which received public funds and protected pupils with government assisted places in private schools.

57 *Costello-Roberts* v *United Kingdom* (1994) 19 EHRR 112. Although the European Court found by five to four votes that a 'slippering' administered by a headmaster had not infringed a school-boy's

© The Modern Law Review Limited 1999

1993[58] warning independent schools that they could not lawfully administer corporal punishment which could be described as being 'inhuman or degrading'. In other words, they were directed to heed the terms of Article 3 regarding the punishment of their pupils. The School Standards and Framework Act 1998 finally outlawed the use of corporal punishment in all schools, state and independent alike. Indeed, it goes much further than previous legislation in so far as it makes the use of corporal punishment in schools unlawful not only for civil purposes but also for criminal purposes.[59]

Despite the Convention being utilised to attack the legality of corporal punishment in schools, it had not hitherto been used by children as a means of challenging their *parents'* use of corporal punishment in the home. It was recently established, however, that Article 3 can protect children against such parental behaviour.[60] A young British boy successfully argued that his step-father's ability to beat him with a garden cane at intervals over a period of a week, as a form of punishment, without any criminal sanction, infringed his rights under Article 3.[61] The European Court confirmed the European Commission's unanimous decision in the boy's favour.[62] Like the Commission, it considered that the United Kingdom government must take responsibility for providing children with practical and effective state protection against treatment or punishment contrary to Article 3.[63] This it had manifestly failed to do since a jury had been able to acquit the stepfather of behaviour of a severity clearly prohibited by Article 3. Both Commission and Court emphasised their view that the English common law defence to a charge of assault against a child, namely that the parent's behaviour amounted to 'reasonable chastisement',[64] undermined the law's ability to protect children's rights. As in other recent case law,[65] both bodies attached importance to the way in which the UN Convention on the Rights of the Child protects children's rights in this context. Indeed, the Commission discussed at some length[66] the criticisms levelled by the Committee on the Rights of the Child at British law on this topic.[67]

The impact of the recent decision was restricted by the European Court's failure to suggest that corporal punishment *per se* amounts to a violation of a child's rights under Article 3. Indeed, it reasserted its earlier view that it depends on the circumstances of each case whether ill-treatment reaches the level of severity required to infringe Article 3.[68] Unfortunately this allows the decision to be downplayed by arguing that it merely precludes parents from treating their children in a particularly brutal way. Thus the Health Minister rapidly responded to it by

rights under Article 3, the Court pointedly voiced (at para 36) severe misgivings about the retention of corporal punishment in British schools.

58 Education Act 1993, later re-enacted as Education Act 1996, ss 548–549

59 See Education Act 1996, s 548–549, as amended by School Standards and Framework Act 1998, s131.

60 *A* v *United Kingdom* n 51 above

61 Neither the European Commission nor Court considered it necessary to determine the boy's claim that his rights under Article 8 had also been violated.

62 *A* v *United Kingdom* (Application No 25599/94) (1996) 22 EHRR CD 190.

63 Both Commission and Court referred to *X and Y* v *The Netherlands* (1986) 8 EHRR 235 (a case involving Article 8) to support their view that positive measures must be taken to provide an effective form of deterrence against violations of Article 3.

64 Children and Young Persons Act 1933, s1(7) and *R* v *Hopley* (1860) 2 F and F 202.

65 eg *Hokkanen* v *Finland*, n 49 above.

66 *A* v *United Kingdom*, n 62 above, para 52

67 *Concluding Observations of the Committee on the Rights of the Child: United Kingdom of Great Britain and Northern Ireland*, CRC/C/15/Add 34 Centre for Human Rights, Geneva, January 1995, para 16.

68 *Costello-Roberts* v *United Kingdom* n 57 above, para 30.

© The Modern Law Review Limited 1999

claiming that the 'overwhelming majority of parents know the difference between smacking and beating'.[69] Nevertheless, the government has accepted that the law requires clarifying to 'reflect the view that violence against children is unacceptable and promote better protection for children without getting in the way of normal family life'.[70]

It seems clear that, at the very least, the parental right to administer lawful chastisement[71] must be revoked and that parents must be warned that discipline amounting to 'inhuman or degrading treatment' is legally prohibited. The legislative attempt at providing comprehensible guidance to schools on the type of punishment falling into such a category[72] suggests that parents will be left in an extremely confused state. Certainly such a change would not bring about the complete abolition of corporal punishment that many consider is required. Experience elsewhere shows that laws of such kind can be extremely effective. Research in Sweden suggests a strong decrease in the use of physical punishment since 1979 when the law there was changed prohibiting any form of smacking children.[73] Nevertheless, to introduce such a change here would be highly controversial and not one that any government would contemplate making with equanimity.

Whether or not the government grasps the nettle on this issue, the Human Rights Act 1998 may eventually provide the means of achieving such a change. This is quite simply because, as discussed above, the courts themselves, being 'public authorities' are obliged to act in a way which is compatible with the Convention.[74] Under the new system, a domestic court may be confronted with a child's claim that the existing legislation, by authorising the use of *any* form of corporal punishment by her parents is incompatible with her rights under Articles 3, 8 and 14. Consistent with an appreciation that the interpretation of the Convention must keep pace with the changing needs of society,[75] the domestic court might consider it appropriate to rule that any form of corporal punishment not only infringes a child's right to be protected from inhuman and degrading treatment under Article 3, but also infringes her right to physical integrity under Article 8[76] and discriminates against her right to receive protection equal to that accorded to adults, thereby infringing her rights under Article 14.[77] It would then be obliged to re-cast the principles of common law in accordance with this interpretation of the Convention. One way or the other, the law on corporal punishment seems destined for reform.

---

69 'Health Minister welcomes "common sense decision"', Department of Health press release, 23 September 1998.

70 *ibid.*

71 See Children and Young Persons Act 1933, s1(7).

72 See Education Act 1996, s 549(3) (now repealed) which contained guidance to schools on the type of corporal punishment which might fall into such a category

73 Whereas in a 1979 survey about half of Swedish parents reported that they spanked their children (9 per cent used spanking every or every other week), a 1996 survey indicated that roughly 30 per cent of middle school pupils and a higher proportion of adults reported having been physically punished by their parents before they reached their teen years. *Spanking and Other Forms of Physical Punishment: A study of adults' and middle school students' opinions, experience and knowledge* Demography, the Family and Children (Stockholm: Statistics Sweden (SCB), 1996) 15.

74 HRA 1998, s 6 (1) and (3).

75 *Cossey* v *United Kingdom* [1991] 2 FLR 492, 501.

76 Article 8 may provide a much more effective means of protecting children against corporal punishment than the rigid requirements of Article 3. See the concurring opinion of Mr E A. Alkema in *A* v *United Kingdom* (Application No 25599/94) n 62 above.

77 In *A* v *United Kingdom* n 51 above, the European Court did not find it necessary to rule on the applicant's claim under Article 8 and his claim under Article 14 was withdrawn.

© The Modern Law Review Limited 1999

Right to information about family origins

The psychological importance of children gaining information about their family origins is now widely acknowledged. In some circumstances, they may be assisted in their search to find out about their early lives by the requirements of Article 8. In *Gaskin*, the European Court ruled that the concept of respect for a person's private and family life is a very broad one and protects a person's vital interest to obtain information necessary to know and understand his childhood and early development.[78] Gaskin successfully argued that Liverpool City Council's refusal to allow him access to a complete set of records relating to his childhood whilst in local authority care amounted to a breach of Article 8. In fact, by the time the European Court considered his application, legislation and regulations had already established the general principle that people should be able to discover what is said about them in local authority records. Not only was such a development too late for Gaskin, but there were important qualifications to that general principle.[79] The European Court accepted the importance of maintaining confidentiality regarding public records, but it did not consider that the system whereby access to records was dependent on the contributor's consent was in conformity with the principle of proportionality, in the absence of an independent authority to arbitrate on the reasonableness of the contributor's refusal to agree to disclosure.[80] The Court's decision confirmed that the newly established principle of openness in local government was of immense importance, especially for the children who spend many years of their childhood cared for by local authorities.[81] The Access to Health Records Act 1990 soon pushed the right to access manually compiled personal information into the medical domain.[82] Although under new legislation the government intends to preserve existing exemptions to full disclosure of personal information,[83] it appears that the decision in *Gaskin* has at last been fully implemented in this regard. For example, the reasonableness of a decision to withhold information because it will cause the individual seeking its disclosure serious harm to his health, could now be scrutinised objectively by the Data

78 *Gaskin* v *United Kingdom* (1990) 12 EHRR 36.

79 LAC (83)14 (made pursuant to Local Authority Social Services Act 1970) set out the principles governing disclosure of social work records but para 5 exempted information supplied in confidence if its authors had not consented to its disclosure. Because only 19 of the 46 contributors had agreed to their disclosure, only 65 of the 352 documents held on Gaskin's file were released to him. See also the Access to Personal Files Act 1987 (replaced by the Data Protection Act 1998) which imposed a duty on local authorities to provide full access to social work records. But see Access to Personal Files (Social Services) Regulations SI 1989/206, reg 5, 8(4) and 9, all of which exempt information from disclosure in certain circumstances. For a general discussion of this area of law, see P. Birkinshaw, *Freedom of Information: The Law, the Practice and the Ideal* (London. Butterworths, 1996) ch 6.

80 *Gaskin* v *United Kingdom* n 78 above, para 49.

81 Department of Health guidance LAC (89)2 indicates that children sufficiently mature to understand the nature of the request may apply for access to social work records; for children not deemed to fall into this category, parents may apply on their behalf. See R Pickford, 'The Gaskin case and confidential personal information – whose secret?' (1992) 4 JCL 33.

82 But access may be withheld from the patient in certain circumstances, including where the information might cause serious harm to the physical or mental health of the subject or identify the informant. See Access to Health Records Act 1990, s 5 and Access to Health Records (Control of Access) Regulations 1993, SI 1993/746

83 The Data Protection Act 1998 s 30 provides the Secretary of State with a power to exempt by order certain types of personal information, such as information relating to health and social work (s30(1) and (3)) from the general provisions regarding right of access to personal data. It appears that the existing exemptions regarding information originating from a health professional or used in social work practice which might cause serious harm to the physical or mental health of the subject or breach the confidences of third parties will be preserved by order of the Secretary of State. See *Data Protection Act 1998: Consultation Paper on Subordinate Legislation*, Home Office (1998), para 26.

© The Modern Law Review Limited 1999

Protection Commissioner,[84] thereby complying with the advice of the European Court of Human Rights.

Gaskin's success in the European Court has led to the suggestion that current legislation which prevents children conceived by artificial reproduction gaining information about the identity of their birth parents might similarly be challenged as being incompatible with their rights under Article 8 of the European Convention.[85] Were such a claim to be brought, the government would almost certainly argue that disclosure of identifying information about donors of gametes and embryos would deter donors and undermine infertility clinics' ability to provide an effective treatment service.[86]

Right to life

In some cases there may be a family dispute over the child's right to life itself. Convention case law suggests that Article 2[87] only fully protects the lives of people who have been born and that therefore claims cannot be brought thereunder to protect an unborn foetus against abortion.[88] Lansdown has suggested that the distinction made by the abortion legislation between the disabled foetus and those considered to be without disabilities is discriminatory and therefore in breach of Article 14.[89] Although ingenious, this argument seems difficult to substantiate, particularly as Article 14 does not normally operate independently of other Articles and since the only Article it could conceivably be linked to is Article 2, which as commented above, has been ruled out by earlier case law.

There is a greater chance of Article 2 being used in the domestic courts to underwrite a claim of incompatibility in circumstances similar to those arising in *R* v *Cambridge District Health Authority, ex parte B.*[90] A child suffering from terminal leukaemia[91] argued unsuccessfully that the Cambridge District Health Authority had behaved unreasonably in refusing to fund treatment which might give her a chance of life. The approach of Laws J, at first instance, foreshadowed that required from the judiciary after incorporation. He noted that there are certain fundamental rights, including the right to life, as recognised by Article 2, so important that they should now be recognised as part of the common law. The health authority had not satisfied him that there was a sufficient justification for refusing the child her chance of life. The Court of Appeal reversed his decision since they did not consider the health authority's decision to be 'unreasonable' for judicial review purposes. Nevertheless, the judgment of Laws J might usefully inform future claims which focus not on judicial review unreasonableness but on incompatibility with Convention rights. Such a claim might however be countered

84 Data Protection Act 1998, s 42.

85 See G. Van Beuren, n 19 above, 33.

86 And was therefore justified under Article 8(2), as being 'necessary in a democratic society .... for the protection of health or morals, or for the protection of the rights and freedoms of others'.

87 'Everyone's right to life shall be protected by law'.

88 See especially *Paton* v *United Kingdom* (1980) 3 EHRR 408. See a more detailed discussion in D. Feldman, *Civil Liberties and Human Rights in England and Wales* (Oxford. Clarendon Press, 1993) 105–107.

89 See G. Lansdown, 'European Convention on Human Rights: Implications for Children' (1997) unpublished paper. The Abortion Act 1967 protects a foetus without abnormalities from abortion once it has attained 24 weeks development By contrast an abortion can be carried out on a disabled foetus right up to term.

90 [1995] 1 FLR 1055. See the discussion by J. Fortin, n 23 above, 255–256 See also J. Fortin, '*Re C (medical treatment)*' (1998) 10 CFLQ 411 for a discussion of a more recent decision involving not dissimilar issues

91 She was later identified as Jamee Bowen.

© The Modern Law Review Limited 1999

by health authorities urging that whilst Article 2 protects individuals against being killed, it imposes no positive obligations on public authorities to devote resources to extending a patient's life without balancing competing public interests. In any case, to prioritise spending using a patient's chances of survival as an overriding criterion might not necessarily promote his or her well-being or produce an effective health service. It will be interesting to see the extent to which the domestic courts allow such arguments to be balanced against Article 2 claims that the right to life must not be infringed, at any cost.

Children's educational rights

One of the few parts of the Convention to secure a right of specific importance to children, as opposed to adults, is Article 2 of the First Protocol which proclaims that

> No person shall be denied the right to education. In the exercise of any functions which it assumes in relation to education and to teaching, the State shall respect the right of parents to ensure such education and teaching in conformity with their own religious and philosophical convictions.

Nevertheless, by simply protecting the right to education *per se*, this provision has none of the fine-tuning required to accommodate the full range of children's educational needs. Our domestic education law goes further and acknowledges that the type of educational provision suitable for one child may be useless to another. Local education authorities are not only under an obligation to provide all school-aged children with free education, but education which is appropriate to their 'different ages, abilities and aptitudes'. [92] Children and parents dissatisfied with the type of educational provision offered might now be tempted to reinforce their complaints by employing the admittedly tenuous argument that the provision of inappropriate educational provision is not only in breach of domestic education law but also infringes a child's rights under Article 2 of the First Protocol. The local education authority might counter such a claim by arguing that it has a measure of discretion in the way it fulfils its duties to provide a child with education.

There are probably more reliable ways in which incorporation may bring about improvements in the existing principles of education law. As many have observed, the legislation in this field appears to ignore children's ability to reach sensible decisions of their own regarding their education. Throughout the reforming activity of the 1980s and early 90s,[93] the education legislation consistently treated parents as the 'consumers' of education, and as their children's proxies or representatives when making educational choices.[94] This established a very different system to that envisaged by those involved in the children's rights movement in English schools in the late 1960s and early 70s when student militancy spread from universities into the schools.[95] The Schools' Action Union (SAU), the National Union of School Students and other bodies[96] asserted a range of pupils' rights in schools. They wanted to establish educational democracy through school councils, the abolition

92 Education Act 1996, s14 (3)

93 See inter alia. Education Acts 1980 and 1981, Education (No 2) Act 1986, Education Reform Act 1988, Education (Schools) Act 1992, and Education Act 1993.

94 See N. Harris, *Law and Education: Regulation, Consumerism and the Education System* (London: Sweet & Maxwell, 1993) 19–22.

95 See C. Wringe, *Children's Rights: A Philosophical Study* (London: Routledge and Kegan Paul, 1981) ch 1 for a description of the 'pupils' rights movement' in schools.

96 eg the National Council for Civil Liberties who in 1970–71 published a series of discussion papers on children's rights and the Advisory Centre for Education (ACE) who in 1971 published a draft Charter of Children's Rights.

© The Modern Law Review Limited 1999

of school uniform and corporal punishment and the right to freedom of expression. The SAU demanded pupils' right to publish in school magazines without censorship, to organise student meetings at lunch break and after school on school premises, to join student unions and engage in political activity including strikes.[97] As discussed elsewhere,[98] it is unclear why this movement ran out of steam so quickly. The explanation may be that unlike countries with a written constitution, the common law system here hampered children's advisers when considering what legal principles might promote the type of rights they were claiming in their schools. Although the judicial review procedure can, as demonstrated by *ex parte B*[99] discussed above, be employed to challenge infringements of human rights, claims quickly founder on technicalities such as judicial review unreasonableness, rather than allowing the courts to discuss more fundamental values.

At least on the face of it, the wording of the Convention appears to provide little scope for challenging the principles of domestic education law which give educational choices to parents rather than to children. In particular Article 2 of the First Protocol seems to reinforce them by directing States to 'respect the right of *parents* to ensure such education and teaching in conformity with their own religious and philosophical convictions'.[100] This provision was clearly designed to protect parents from religious and racial persecution by the state indoctrinating their children through their schooling. It should be pointed out at this juncture that the government is intent on avoiding minority ethnic groups arguing that their rights under the Convention are infringed by a refusal to fund the establishment of denominational schools to cater specially for their children. The United Kingdom has maintained its existing reservation to the second sentence in Article 2 of Protocol One, indicating that it accepts this principle 'only so far as it is compatible with the provision of efficient instruction and training, and the avoidance of unreasonable public expenditure'.

Not surprisingly, given its wording, Article 2 of the First Protocol has led to a variety of claims which, despite involving aspects of children's education, are in reality complaints about infringements of parents' own strong philosophical convictions.[101] Indeed, ironically, its terms have been exploited to undermine children's rights to independent thought. For example, the United Kingdom government argued[102] that they justified retaining parents' rights to withdraw their children from the religious instruction and collective school worship provided in state schools[103] and from sex education classes.[104] This argument was unfounded; the European Court had in fact decided that compulsory sex education in schools does not violate parents' rights under that Article as long as it is conveyed in a balanced and objective manner, with no element of indoctrination.[105] But under current legislation British parents are neither required to provide any reason for exercising their right of withdrawal nor to consult their children over such a decision. Domestic law thereby reinforces parents' perceptions that they have a right

97 In 1972 approximately 2,500 pupils absented themselves from school to attend a 'Schools Demo' in Trafalgar Square mounted by the SAU.

98 See J Fortin, n 23 above, 132 et seq.

99 *R* v *Cambridge District Health Authority, ex parte B* [1995] 1 FLR 1055.

100 Emphasis added.

101 eg *Kjeldsen, Busk Madsen and Pedersen* v *Denmark* (1976) 1 EHRR 711.

102 See Baroness Blatch, HL Deb vol 547 col 1292 6 July 1993.

103 See Education Act 1996, s 389.

104 *ibid* s 405.

105 *Kjeldsen, Busk Madsen and Pedersen* v *Denmark* n 101 above, where Danish parents failed to establish that the Danish government had violated their own rights under that provision by providing compulsory sex education in schools See J. Fortin, n 23 above, 150–155.

© The Modern Law Review Limited 1999

to dictate their views on all these matters to their children. After incorporation of the European Convention, such provisions might be challenged by an adolescent claiming that whilst these statutory exemptions comply with her parents' rights under Article 2 of the First Protocol of the Convention, her own rights to freedom of expression, of religion and to receive and impart information are all being infringed.

In the past, complaints to the European Commission and Court about children's rights to education have often been brought by parents and children together, linked with claims to religious freedom and freedom from discrimination. For example, in *Angeleni* v *Sweden*[106] a mother and her daughter claimed infringement of their rights under Articles 9, 14 and 17. The Swedish Ministry of Education had refused to exempt the daughter from religious education on the basis that as an atheist she was not a member of an organised religion. Since the daughter was only eight years old however, it is questionable to what extent her views were dictated by her mother. Although parents' attempts to protect children from religious indoctrination by the state are perfectly proper, cases like this provoke the uneasy feeling that children's claims may sometimes stem from *parental* indoctrination.

To date, the European case law appears to provide little ammunition for those wishing to strengthen children's rights to make their own educational choices rather than allowing their parents to do so for them. There may, however, be grounds for more optimism regarding other aspects of education law. Many have criticised those provisions in the education legislation which exclude children under 18 from being parties to appeals heard by education tribunals.[107] The children most affected by this aspect of the existing law are those who are permanently excluded from school and those with special educational needs. Giving parents the sole right of appeal to a tribunal over such matters makes a false assumption that children's interests are always identical with those of their parents. A child's inability to become involved in the appeal processes deprives him or her of an important procedural right. This is particularly unjust in cases where, for example, parents are too disinterested or nervous to pursue an appeal on the child's behalf, or where they are incapable of presenting his or her case satisfactorily or objectively. A child might therefore persuade a domestic court that English education law is infringing his or her right to a 'fair and public hearing' under Article 6 and that Strasbourg's early and very narrow interpretation of that Article's scope is now anachronistic.[108] A declaration of incompatibility would force the government to review the present system of appeals. Alternatively, a court confronted with a challenge from an adolescent seeking party status in an appeal hearing might simply comply with its duty to interpret the legislation 'so far as it is possible to do so', in accordance with the terms of Article 6.[109] It might therefore decide that since only a 'relevant person' in the current

106 (1988) 10 EHRR 123.

107 Education Act 1996, Sched 16, para 17 and Schools Standards and Framework Act 1998, s 65(5) and Sched 18 para 17.

108 Early convention case law interpreted the phrase 'civil rights and obligations' in Art 6 to mean only those rights pertaining to private rather than public law. Consistent with this approach in *Simpson* v *UK* (1989) (Application No 14688/89) ECHR 1989–90 Vols 62–64 Decision and Reports 188 the European Commission refused to consider a mother's claim that the procedures used to determine her son's special educational needs were in breach of Art 6. This was because his right to education was not of a civil nature, falling as it did within the domain of public law. But see D. Harris, M. O'Boyle and C. Warbrick, *Law of the European Convention on Human Rights* (London: Butterworths, 1995) 180–186 for a discussion of the later Strasbourg authorities which reflect a more flexible approach to the public/private law distinction. These authors suggest that the European Court should reformulate its approach to interpreting the phrase 'civil rights and obligations'.

109 Thereby complying with its duties under HRA 1998, s 3 (1) to interpret primary legislation in a way which is compatible with Convention rights.

© The Modern Law Review Limited 1999

education legislation can have party status, the term 'relevant person' must be interpreted to include children under the age of 18, despite the obvious contrary intention of the legislative definition.[110]

Young offenders

The present Home Secretary is determined to 'tackle' youth crime. Like his predecessors, he shows little real sympathy for the view that, unlike adults, many young people are developmentally incapable of taking responsibility for their behaviour. Indeed, he considers that the present youth justice system 'too often excuses the young offenders before it, implying that they cannot help their behaviour because of their social circumstances'.[111] The Crime and Disorder Act 1998 aims to establish a system whereby young offenders will be 'confronted with their behaviour and helped to take more personal responsibility for their actions'.[112] But a sophisticated legal system should surely accommodate the fact that young offenders are very different to adult perpetrators. They are often the product of their upbringing and may with education and support, mature and grow out of their offending behaviour. Similarly it should acknowledge that young offenders do not, by virtue of their offending behaviour, forfeit all those rights that other citizens take for granted. Despite this, successive governments have adopted increasingly harsh attitudes when formulating legislation for young offenders.[113]

There has been considerable criticism of various aspects of the Crime and Disorder Act 1998.[114] The intention to clamp down on any person over the age of 10 behaving in a manner which caused or 'was likely to cause harassment, alarm or distress to one or more persons not of the same household as himself . . .' through the imposition of 'anti-social behaviour' orders[115] hardly reflects a refreshingly liberal approach to the proper role of the state. This, linked with the notion that children under the age of 10 should be kept in their homes during long summer evenings by local child curfew schemes,[116] smacks of extreme authoritarianism. The undermining of fundamental rights can be an insidious process, particularly if it is masked by appeals to middle class conformist sentiments. Unfortunately, incorporation of the European Convention may not achieve a more generally humane approach to dealing with young offenders. As a group, they receive greater protection from more child focused documents such as the UN Convention on the Rights of the Child, which contains very detailed provisions regarding their arrest and treatment.[117] Some of the Articles listed in the European Convention have, however, provided useful pressure for reform and doubtless they will continue to do so. For example, Article 5 has been used to improve the protection available for some young offenders, although only on a case by case basis and in a rather ad hoc way. Following a European Court ruling that the United Kingdom was in breach of

110 The term 'relevant person' means in relation to a child under 18, that child's parent. See n 107 above, and *S (a minor)* v *Special Educational Needs Tribunal* [1996] 2 All ER 286.

111 See *No More Excuses – A New Approach to Tackling Youth Crime in England and Wales*, Cm 3809 (London: HMSO, 1997) preface.

112 *ibid.*

113 The approach of successive governments to the rights of young offenders is discussed generally by J. Fortin, n 23 above, ch 19.

114 See (1997) *Criminal Law Review* 769 and 848 and (1998) *Criminal Law Review* 1 See also A. Rutherford, 'A Bill to be Tough on Crime' (1998) 148 NLJ 13.

115 Crime and Disorder Act 1998, s 1

116 *ibid* s 14.

117 Articles 39 and 40. See G Van Beuren, n 19 above, ch 8 and Fortin, n 23 above, ch 19.

© The Modern Law Review Limited 1999

Article 5(4),[118] the government was obliged to introduce legislation improving the procedure by which young offenders kept in detention under indeterminate sentences gained their eventual release after expiry of the tariff period.[119] The United Kingdom changed its law by transferring responsibility for deciding on the timing of the final release of offenders such as the killers of James Bulger from the Home Secretary to the Parole Board.

Article 6 may provide the means for challenging the method of formal trial adopted for very young offenders charged with serious offences. Thompson and Venables, the 10 year-old killers of James Bulger, have petitioned Strasbourg arguing that they lacked the capacity to cope adequately with their public trial by jury in a Crown Court. They were neither able to understand fully the procedure nor instruct their defence counsel.[120] Future challenges in the domestic courts might argue that only a trial adapted to take special account of children's lack of maturity would fully accord with the aims of Article 6. Whilst such claims have obvious merit, any reforms must also take account of the restrictions imposed by the need for procedural fairness. For example, the European Court has stressed that whatever the special nature of the adjudication, there is, as a matter of general principle, a right to a fair adversarial trial, together with its attendant requirements, such as a prior right to disclosure of all evidence to be called.[121] Thus although the system of children's hearings in Scotland narrowly escaped a detailed scrutiny by the European Court in *McMichael* v *United Kingdom*,[122] the decision flagged up a warning that its procedures required some adjustment.[123] The case also reinforces the need for caution before attempting to make procedures for dealing with children in England and Wales less formal and confrontational. Those advocating a 'welfare approach' to juvenile offending must take full account of the need for procedural fairness and the limitations that this may impose.[124]

The dangers of corrupting young offenders by accommodating them with adult criminals was recognised by the leaders of the nineteenth century reformatory movement. Today, although it is widely acknowledged that the law should ensure that young offenders are kept separate from adult offenders whilst in detention, there continues to be insufficient suitable accommodation for young people entering the criminal justice system. The current system operating in the United Kingdom whereby some juveniles are held on remand in adult prisons has therefore been widely condemned. Through this objectionable practice, vulnerable young people are subjected to the violent and intimidating behaviour that is common in such establishments, a situation which too often leads to suicide attempts or to self-harm. It is arguable that such treatment amounts to inhuman or degrading treatment or punishment and is therefore incompatible with Article 3.[125] Such arguments would be unnecessary were the United Kingdom to abide by the requirements of Article 37(c) of the UN Convention on the Rights of the Child

118 *Hussain* v *United Kingdom* (1996) 22 EHRR 1

119 Crime (Sentences) Act 1997, s28.

120 The applications made on behalf of Thompson and Venables claiming inter alia that the form criminal trial infringed their rights under Articles 3 and 6 have been declared admissible by the European Commission of Human Rights. *T* v *United Kingdom* (Application No 24724/94) (1998) 25 EHRR CD 11 and *V* v *United Kingdom* (Application No 24888/94) (1998).

121 *McMichael* v *United Kingdom* (1995) 20 EHRR 205.

122 *ibid.*

123 The European Court implicitly disapproved of the fact that parties to the hearing did not routinely receive copies of all reports and documents relevant to the proceedings either before, during or after the hearing; *ibid* para 80.

124 See J. Fortin, n 23 above, 448–449.

125 See G. Lansdown, n 89 above.

© The Modern Law Review Limited 1999

which clearly requires 'every child deprived of liberty to be separated from adults ...'. Unfortunately, when the United Kingdom ratified the UN Convention, it entered a reservation to that Article entitling it to exemption where the lack of facilities prevent compliance.

The criminal law must take care not to criminalise the behaviour of one group of children and young people, whilst leaving another group free to behave in a similar fashion without interference. This was made clear when the European Commission upheld Sutherland's claim that English law violates the rights of homosexual males between the ages of 16 and 18 under Articles 8 and 14 of the Convention.[126] By criminalising their sexual activities, the law not only fails to respect their private lives[127] but also discriminates against them by allowing lesbians and heterosexual young people to consent to sexual relations at the age of 16.[128] To forestall the European Court of Human Rights confirming the Commission's decision, the United Kingdom government lodged with the European Court an undertaking to take the earliest opportunity to introduce a free vote on the equalisation of the age of consent. Although recent legislative efforts to achieve such a change have so far failed, it appears that the promised reform is imminent.[129]

Conclusion

The discussion above illustrates the strengths and weaknesses of the European Convention when applied to children. The list of rights it contains and the Strasbourg jurisprudence surrounding their interpretation has affected children in a variety of unpredictable ways. At best, there are the relatively rare applications taken to Strasbourg, brought directly on behalf of children, such as those relating to the use of corporal punishment, which produce recognition and enforcement of fundamentally important rights. Next are the vast majority of claims brought by adults, on their own behalf, which may benefit children incidentally, since as human beings, they too are the recipients of Convention rights. Third, the Convention can at times operate in a merely neutral way regarding children, in so far as the rights it protects do not cater for their particular needs, but nor do they actually damage children in their application. This is not to say that children do not value the civil and political freedoms the Convention protects, merely that since children's lives are largely led in the protected environment of their homes and schools, there are not many obvious opportunities for infringing such rights.

Lastly, and most worryingly, the European Convention may at times actually undermine the promotion of children's rights. This is because the Articles and Convention case law reinforce the common view advocated by most legal systems throughout the world that the value of family life lies in its privacy from state

126 *Sutherland* v *United Kingdom* (Application No 25186/94)(1997) 24 EHRR CD 22.

127 Respect for an individual's private life under Article 8 includes respect for his sexual life.

128 Sexual Offences Act 1956 s14 (2) prevents a girl under the age of 16 consenting to any act which would be an indecent assault. Sexual Offences Act 1967 s 1 (as amended) provides that a male who has attained the age of 18 may consent to a homosexual act in private.

129 In the summer of 1998, the Crime and Disorder Bill was amended by introducing a new cl 1 lowering the age of consent to homosexual sexual relations to the age of 16. Although a majority voted in favour of this amendment in the House of Commons, it failed to obtain House of Lords approval. See R. Swade, 'The Crime and Disorder Bill: The Age of Consent Debate' (1998) 148 *Childright* 7. The government has persevered with its efforts and reintroduced amending legislation in the form of the Sexual Offences (Amendment) Bill. The House of Commons, in January 1999, voted in its favour but peers may renew their attempts to block such a change when the Bill reaches the House of Lords.

© The Modern Law Review Limited 1999

interference; furthermore, that family privacy involves parental autonomy.[130] Whilst Article 8 of the Convention carries this message most clearly, the other Articles provide little encouragement to view children as individuals, with their own rights, independent from those of their parents and other family members. The willingness of the European Court of Human Rights to assist parents, has demonstrated a worrying inability to perceive that such an approach may directly damage children's rights. Furthermore it sets a dangerous precedent for the domestic courts when faced with similar adult claims. Indeed, there is a considerable risk that incorporation of the Convention may encourage the increasingly laissez-faire approach adopted by English law regarding the regulation of family life. As discussed elsewhere, the Children Act 1989 'not only promoted the assumption that responsible parents would automatically promote their children's interests without legislative encouragement, but was also intent on reinforcing the privatisation of family life, by withdrawing the law from areas where it had previously had some influence ...'.[131] Successive governments in this country have shown a marked reluctance to offend public opinion on such matters as state interference with parental freedom. There is now the enticing prospect of exploiting the narrowness of the European Convention to justify their refusal to champion children's rights.

On a more optimistic note, incorporation may at the very least produce a different type of jurisprudence which casts the principles of law applying to children in terms of their human rights, such as freedom of expression and assembly, rather than being restricted to the more mundane principles of common law. The existence of a list of rights available to adults and children alike should allow the boundaries to children's rights to be tested in the domestic courts, as they have been in the USA. For example, the American Supreme Court in *Tinker* v *Des Moines Independent Community School District*[132] ruled that 'Students in school as well as out of school are "persons" under our Constitution. They are possessed of fundamental rights which the State must respect ...'.[133] The extent to which our domestic courts consider children to be 'persons' and therefore entitled to the same rights as adults under our new Human Rights Act will, on implementation of the legislation, begin to emerge.

130 See Fortin, n 23 above, 225–228.
131 *ibid* 228.
132 393 US 503 [1969]. See also *Goss* v *Lopez* 419 US 565 [1975].
133 *ibid* 511.

© The Modern Law Review Limited 1999

Nationalism and the Pathology of Legal Systems: Considering the *Quebec Secession Reference* and its Lessons for the United Kingdom

Mark D. Walters*

Questions relating to the disintegration and destruction of legal systems – or what H.L.A. Hart called the 'pathology of legal systems'[1] – are extremely difficult to confront as questions of law. When political allegiance to one legal system begins to erode, the natural tendency of judges may be to struggle to save the system from collapse, thereby putting at risk the chance of orderly transition to succeeding nascent legal systems. However, just as it is best for a natural person to order his or her affairs and write a will while of sound mind, perhaps a legal system – even one with promise and potential remaining – should likewise plan for its orderly, and lawful, demise. Indeed, by confronting the unthinkable, perhaps it can (unlike a natural person contemplating eventual death) achieve partial survival, so that in the aftermath of its dismemberment a remnant of the system remains worthy of allegiance from at least some of its citizens.

This was, in essence, the task that the Supreme Court of Canada accepted in *Reference re the Secession of Quebec from Canada*.[2] In what the Chief Justice of Canada, Antonio Lamer, described as the most important judgment in the Court's history,[3] the nine Supreme Court justices ruled unanimously that the Dominion of Canada is divisible and (by implication) destructible, and that there are *legal* rules governing such division and destruction. The case arose when the federal government of Canada invoked the advisory jurisdiction of the Court by referring to it three questions, namely: (1) Can the government or legislature of the province of Quebec effect the secession of Quebec from Canada unilaterally under Canadian law? (2) Can the government or legislature of Quebec effect secession unilaterally under international law? (3) If Canadian and international law conflict on this point, which prevails?[4] Thus, the Court confronted the same basic legal issue as did the Judicial Committee of the Privy Council in the famous case of *Madzimbamuto* v *Lardner-Burke*[5] with the critical difference that, in that case, the Rhodesian government had

* New College, Oxford

I am grateful to Peter Oliver for his comments on an earlier draft of this article

1 H.L.A. Hart, *The Concept of Law* (Oxford: Clarendon Press, 1961) 114.

2 *Reference by the Governor in Council, pursuant to s 53 of the Supreme Court Act, concerning the secession of Quebec from Canada* [1998] 2 SCR 217, hereinafter *Quebec Secession Reference*.

3 Tu Thanh Ha, 'Quebec on edge as historic ruling set for Thursday' *The Globe and Mail* (Toronto) 15 August 1998, A1.

4 The full text of the questions is: '1. Under the Constitution of Canada, can the National Assembly, legislature or government of Quebec effect the secession of Quebec from Canada unilaterally? 2. Does international law give the National Assembly, legislature or government of Quebec the right to effect the secession of Quebec from Canada unilaterally? In this regard, is there a right to self-determination under international law that would give the National Assembly, legislature or government of Quebec the right to effect the secession of Quebec from Canada unilaterally? 3. In the event of a conflict between domestic and international law on the right of the National Assembly, legislature or government of Quebec to effect the secession of Quebec from Canada unilaterally, which would take precedence in Canada?' (Order in Council PC 1996–1497 30 September 1996).

5 [1969] 1 AC 645.

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

already made its unilateral declaration of independence (UDI) by the time its legality was judicially tested – the old legal regime was, in practical terms, dead – whereas in the *Quebec Secession Reference* the Court considered the legality of a UDI by Quebec that has not been, and may never be, made, and in the context of a Canadian constitutional order that, it must be said, still has plenty of life in it.

The Court gave negative answers to the first two questions and, therefore, did not need to address the third. The opinion in the *Quebec Secession Reference* contains much of interest to public lawyers in other common law jurisdictions, as well as other multinational states. Indeed, it will be of particular relevance to those who advocate the independence of Scotland, Wales or Northern Ireland from the United Kingdom. The Court ventured well beyond the textual confines of the Canadian constitution and engaged in the theoretical analysis of general principles of constitutional law, including federalism, democracy, sovereignty, the rule of law, the concept of justiciability and rights of minorities. The purpose of this article is to review the events leading to the reference, to summarise the Court's answers to the questions, and then to discuss in more detail some of the general legal, constitutional and jurisprudential issues that the case raises, with particular reference to its significance to questions of nationalism and secession in the United Kingdom.

Events leading to the *Quebec Secession Reference*: the 1995 referendum and the development of 'Plan B'

On 30 October 1995, the 132-year old federation of Canada came precariously close to collapse: in a province-wide referendum the Quebec electorate voted by a narrow margin (50.56 to 49.44 per cent) against the provincial government's plan for the independence of Quebec from Canada.[6] Had the separatist Parti Québécois (PQ) government won the referendum, it would have initiated a prepared independence strategy: draft legislation provided that Quebec would have offered Canada an opportunity to negotiate a new Canada-Quebec economic and political partnership, but if agreement was not reached within one year or if Canada refused to negotiate, the Quebec National Assembly could declare Quebec to be a 'sovereign country' and seek international recognition as such.[7] As Lesage J observed in *Bertrand* v *Quebec (Procureur General)*, in which an injunction to stop the referendum was sought and denied, the PQ government had undertaken to 'proceed with a unilateral declaration of independence' and was 'seeking to overthrow the constitutional order'.[8]

Prior to the 1995 referendum, Canada's federal government insisted that a separatist referendum victory was unlikely and there was no need to debate or challenge the legality of PQ plans for independence. Shocked by the closeness of

6 The referendum question was: 'Do you agree that Quebec should become sovereign, after having made a formal offer to Canada for a new Economic and Political Partnership, within the scope of the Bill respecting the Future of Quebec and of the agreement signed on 12 June 1995?'. See K. McRoberts, *Misconceiving Canada: The Struggle for National Unity* (Toronto Oxford University Press, 1997) 230–231.

7 An Act Respecting the Sovereignty of Quebec, Quebec, National Assembly, 1st sess., 35th leg (draft bill issued on 6 December 1994 but not enacted) and An Act respecting the Future of Quebec , Quebec National Assembly, 1st sess, 35th leg Bill 1 (introduced in the provincial legislature on 7 September 1995 but not enacted). See, José Woehrling, 'Les aspects juridique d'une éventuelle sécession de Québec' (1995) 74 Can Bar Rev 293, 297–309 and H. Wade MacLauchlan, 'Accounting for Democracy and the Rule of Law in the Quebec Secession Reference' (1997) 76 Can Bar Rev 155, 156–160.

8 *Bertrand* v *Quebec (Procureur General)* (1995) 127 DLR (4th) 408 (Que Sup Ct) 428–429

© The Modern Law Review Limited 1999

the referendum result, however, the federal government revised its policy on Quebec and adopted two parallel plans – 'Plan A' and 'Plan B'[9] – designed to ensure defeat of the separatist cause in any subsequent referendum. Plan A involves convincing Quebec that its distinct cultural and legal heritage can flourish within the Canadian federation, whereas Plan B involves the articulation by the federal government of a clear position on the political and legal consequences of a separatist referendum victory. It has stated that it will not use force to frustrate the democratic will of Quebecers, but that a UDI by Quebec would violate both Canadian and international law.[10] In furtherance of 'Plan B', the federal government referred the question of the legality of a Quebec UDI directly to the Supreme Court of Canada. Thus, the Court was, yet again, drawn into the constitutional-political drama that has been played out in Canada in one form or another since the British conquest of New France in 1759–60.

Before turning to the Court's opinion, it is important to state briefly the nature of the Quebec separatist movement. Canada is a multinational state, and the Quebec separatist movement is (largely) a nationalist movement. It seeks international statehood for one of Canada's nations, or 'peoples', *les Québécois*.[11] Separatists argue that this nation includes all people within the boundaries of the province of Quebec, including the English-speaking community and Aboriginal nations that oppose secession.[12] In response, it is said that *Québécois* national identity exists within the French-speaking community that is composed mainly (but not exclusively) of descendants of settlers of New France, and that other peoples within the province having their own national identities, especially Aboriginal nations, are excluded.[13] Defining the *Québécois* nation and defining the territorial boundaries of an independent Quebec would represent highly complex and emotional aspects of any future negotiation of Quebec's secession.

Debates about national identity, political boundaries and constitutional and legal structure began in Canada soon after Britain conquered New France. Many of the key dates in Canada's constitutional history mark efforts by local colonial and British imperial authorities to accommodate the often conflicting demands of English, French and Aboriginal peoples. In 1763 Canada was divided into an 'Indian' (Aboriginal) homeland and a colony called Quebec populated by a French-Canadian majority and small English-speaking (British) minority.[14] In 1774, Aboriginal nationality was compromised while French-Canadian nationality was recognised: the Quebec Act extended Quebec's borders over much of Indian territory and guaranteed French civil law and religious freedoms in the province.[15]

9 eg P Mohahan, M. Bryant and N Coté, *Coming to Terms with Plan B. Ten Principles Governing Secession* (Toronto. C.D. Howe Institute, 1996).

10 Minister of Justice, Allan Rock, statement to federal House of Commons, 26 September 1996, quoted by McRoberts, n 6 above, 241

11 A. Cassese, *Self-Determination of Peoples· A Legal Reappraisal* (Cambridge: Cambridge University Press, 1995) 248–254.

12 R. Bothwell, *Canada and Quebec: One Country, Two Histories* (Vancouver: UBC Press, 1998) 236–237.

13 D.M. Thomas, *Whistling Past the Graveyard: Constitutional Abeyances, Quebec and the Future of Canada* (Oxford: Oxford University Press, 1997) 224–225; Grand Council of the Crees, *Sovereign Injustice: Forcible Inclusion of the James Bay Crees and Cree Territory into a Sovereign Quebec* (Nemaska, Eeyou Astchee, 1995) ch 1; Grand Council of the Crees (Eeyou Astchee), *Never Without Consent: James Bay Crees' Stand Against Forcible Inclusion into an Independent Quebec* (Toronto: ECW Press, 1998) ch 2.

14 Royal Proclamation of 7th October 1763, printed in A. Shortt and A.G. Doughty (eds), *Documents Relating to the Constitutional History of Canada, 1759–1791* (Ottawa: The Public Archives of Canada; The King's Printer, 1907) 118–123.

15 Quebec Act 1774, 14 Geo III c 83 (GB)

© The Modern Law Review Limited 1999

English-speaking settlers, whose numbers were growing in the 'upper' part of the province, soon insisted on English laws and democratic institutions, and so in 1791 Quebec was divided into two provinces with elected legislative assemblies: Upper Canada, with an English-speaking majority and English common law, and Lower Canada, with a French-speaking majority and French civil law.[16] Though not mentioned in imperial statutes, Aboriginal nations were treated as distinct nations under their own laws and governments within their unceded and reserved territories within the two provinces, though after 1830 concerted efforts were made to assimilate them into settler society and subject them to settler laws.[17]

Conflict between English-speaking executives and French-speaking legislative assemblies erupted in rebellion in the 1830s.[18] In response, the principle of responsible government was introduced into Canada; but to avoid subjecting the English minority in Lower Canada to government by the French-speaking majority, the two provinces were unified as the Province of Canada having an English majority – this would, said Governor General Lord Durham, 'obliterat[e] the nationality of the French Canadians'.[19] A more tolerant spirit soon prevailed in relation to French-Canada, and when the present federation of Canada was constituted by the British North America Act 1867 (the BNA Act; now Constitution Act 1867),[20] the province of Canada was again divided, with the upper, English-speaking province named Ontario and the lower, French-speaking province named Quebec; they were joined in confederation by New Brunswick and Nova Scotia. The process of confederation, then, is often described as a 'compact' between the English and French-speaking nations in British North America – a compact reflected legally, though implicitly, in the selection of a *federal* constitution in which one of the provinces, Quebec, contains a French-speaking majority that can use the extensive powers allocated to provinces as a form of national self-government and self-determination.[21] However, the Quebec province was not intended only as a home for the French-Canadian nation; it had a large English speaking minority and many Aboriginal nations within its boundaries, and in 1898 and 1912 its boundaries were extended north to include Cree and Inuit peoples having distinct national cultures of their own.[22]

Nationalists in Quebec are divided about how the French-Canadian – or *Québécois* – nationality and French language can be better protected and recognised: some opt for outright separation from Canada, others for more explicit recognition of Quebec's distinct status in the Canadian constitution. This ambivalence is reflected in the phrasing of a 1980 referendum question on 'sovereignty-association'.[23] After the defeat of separatists in that referendum, efforts were made to meet Quebec demands for constitutional reform; indeed, meeting these demands became particularly important after the constitution was

16 Constitutional Act 1791, 31 Geo III c 31 (GB); An Act to Introduce the English Law 1792, 32 Geo III c 1 (Upp Can).

17 M. Walters, 'The Extension of Colonial Criminal Jurisdiction over the Aboriginal Peoples of Upper Canada' (1996) 46 U of Toronto LJ 273.

18 Sir R. Coupland, *The Durham Report* (Oxford: Oxford University Press, 1945) 45–58.

19 *ibid* 158. The reunification was achieved under the Union Act 1840, 3 & 4 Vict c 35 (UK).

20 30 & 31 Vict c 3 (UK).

21 Stanley B. Ryerson, *Unequal Union: Roots of Crisis in the Canadas, 1815–1873* (Toronto: Progress Books, 1983) 358–365.

22 An Act respecting the boundaries of Quebec, SC 1898 c 3; An Act to extend the boundaries of the province of Quebec, SC 1912 c 45. See *Sovereign Injustice*, n 13 above, 199–217.

23 On 20 May 1980 the PQ government held a referendum seeking a mandate for the government to negotiate 'sovereignty-association' with the rest of Canada, which agreement would then be put to voters for approval in a second referendum: McRoberts, n 6 above, 156.

 © The Modern Law Review Limited 1999

'patriated' from the United Kingdom and significantly amended in 1982 without Quebec's consent.[24] Included in these amendments was a provision entrenching 'existing aboriginal and treaty rights' – a promise, as the Supreme Court of Canada subsequently held, of a 'just settlement for aboriginal peoples'.[25] Quebec's assent to the patriated constitution, while not legally necessary,[26] remains politically imperative. However, efforts at amending the constitution to recognise, *inter alia*, Quebec as a 'distinct society' failed in the late 1980s and early 1990s.[27] Thereafter the Quebec nationalist sense of betrayal deepened, the separatist movement regained the popularity it had lost after its 1980 referendum defeat, and, in 1995, the PQ government was ready to ask voters again to initiate secession in another referendum – which (as noted) it narrowly lost. Notwithstanding on-going national, cultural and linguistic divisions, the present Canadian constitution has proved to be remarkably resilient since its creation 1867. For judges committed to upholding the constitution's promise and potential, it must have been difficult in the *Secession Reference* to contemplate, even hypothetically, the prospect of its failure and the need to design rules for Canada's dismemberment.

The judgment in the *Quebec Secession Reference*

The Supreme Court of Canada exercises its advisory jurisdiction in much the same way as its adjudicative jurisdiction: it considers oral and written arguments of opposing sides on questions of law (albeit hypothetical ones) and then it issues a reasoned opinion on the state of the law. In the *Quebec Secession Reference*, the Attorney General for Canada argued against the legality of a UDI by Quebec, and the Attorney General for Quebec was entitled to present contrary arguments; however, the PQ government denied the legitimacy of the reference altogether and refused to participate. To ensure that it heard from an opposing side, the Court appointed an *amicus curiae* to present the separatist perspective; it also heard from various interveners, including Aboriginal peoples whose constitutionally-protected aboriginal and treaty rights would be affected by Quebec's secession.

Preliminary objections: the question of justiciability

The *amicus curiae* made a number of preliminary objections, the most significant of which was the argument that the questions referred to the Court were not justiciable because they were 'purely political' and judicial intervention would 'usurp' the democratic decision of the people of Quebec to secede.[28] This argument paralleled statements by the PQ government that the Court had no right to interfere with democracy by telling Quebecers that they could not secede, as well as

24 E. McWhinney, *Canada and the Constitution, 1979–1982* (Toronto: University of Toronto Press, 1982) and K. Banting and R. Simeon (eds), *And No One Cheered: Federalism, Democracy and the Constitution Act* (Toronto: Methuen Publications, 1983).

25 *Sparrow* v *R* [1990] 1 SCR 1075, 1106 (quoting Noel Lyon, 'An Essay on Constitutional Interpretation' (1988) 26 Osgoode Hall LJ 95, 100) Aboriginal and treaty rights are entrenched by the Constitution Act 1982, s 35(1).

26 *Reference re Resolution to Amend the Constitution* [1981] 1 SCR 753 (hereinafter *Patriation Reference*); *Att Gen Quebec* v *Att Gen Canada* [1982] 2 SCR 793; *Att Gen Quebec* v *Quebec Protestant School Boards* [1984] 2 SCR 66.

27 P. Monahan, *Meech Lake: The Inside Story* (Toronto: University of Toronto Press, 1991) and K. McRoberts and P. Monahan (eds), *The Charlottetown Accord, the Referendum, and the Future of Canada* (Toronto: University of Toronto Press, 1993).

28 *Quebec Secession Reference*, n 2 above, para 27.

© The Modern Law Review Limited 1999

arguments made by the Quebec Attorney General in the *Bertrand* case that the question of Quebec's independence was 'extraneous to the rule of law'.[29] Central to the separatist strategy, then, was the propagation of the notion that democracy and the rule of law are, at least in relation to Quebec's status, inherently contradictory, and that democracy must prevail.[30] The Court's response in rejecting the *amicus curiae*'s argument was clearly informed by a desire to defend the legitimacy of the judicial process against allegations – the effects of which were potentially explosive – that it operated in a manner inconsistent with democracy. It therefore held that for it to answer the questions would not be to usurp the democratic decision of the people of Quebec because the questions 'are strictly limited to aspects of the legal framework in which that democratic decision is to be taken', and once this legal framework has been 'clarified', the people of Quebec 'acting through the political process' will be free to exercise their democratic right to decide whether or not to pursue secession.[31] In this remarkable statement, then, the Court acknowledged for the first time that Quebec may *lawfully* secede from Canada, and that the Court would construct the necessary 'legal framework' for secession. The Court was not asked whether Quebec might lawfully secede, let alone what the 'legal framework' for secession is; it was merely asked about the legality of a UDI. Its response to the separatist argument about democracy and the appropriate role of the judiciary was, paradoxically, one of judicial activism aimed ultimately at reconciling (as the Court noted throughout the ruling) 'legality' with (democratic) 'legitimacy'.[32]

Question 1: the legality of a UDI under Canadian law

The Court's response to the first question – would a Quebec UDI be lawful under the Canadian constitution? – was, not surprisingly, 'no'. Indeed, the Court could have reached this conclusion quickly through a literal reading of the constitution's text. The 'Constitution of Canada' includes a number of positive constitutional instruments (listed in section 52(2) and the Schedule of the Constitution Act 1982), none of which expressly permit or prohibit secession of a province or a people from Canada. However, the constitution provides for its own amendment upon ratification of the changes by the federal parliament of Canada and the legislatures of seven provinces having at least 50 per cent of the Canadian population or, in certain specific circumstances, upon ratification by the federal parliament and legislatures of each of the ten provinces; it also permits provinces to amend unilaterally their *own* constitutions.[33] If secession is a change – albeit a radical one – to the Canadian constitution requiring its amendment, then, according to the amendment formulae (whichever applies), no such amendment could be accomplished unilaterally by a seceding province. A UDI by Quebec is therefore unlawful.

The Court adopted this textually-based analysis, but only as a small part of a far larger analysis of the *unwritten* Canadian constitution. Rather than take refuge from the highly politicised aspects of the reference in the safety of positivist legal

29 MacLauchlan, n 7 above, 156, 162–163, *Bertrand* v *Quebec (Attorney General)* (1996) 138 DLR (4th) 481 (Que Sup Ct) 500.

30 R. Howse and A. Malkin, 'Canadians are a Sovereign People: How the Supreme Court Should Approach the Reference on Quebec Secession' (1997) 76 Can Bar Rev 186, 188

31 *Quebec Secession Reference*, n 2 above, para 27.

32 *ibid* paras 67 and 89.

33 Constitution Act 1982, Pt V, ss 38, 41 and 45 (other amendment formulae are provided, but are clearly not applicable to the issue of secession). In general, J R. Hurley, *Amending Canada's Constitution: History, Processes, Problems and Prospects* (Ottawa: Minister of Supply and Services, 1996) ch 4

© The Modern Law Review Limited 1999

analysis, the judges essentially set the constitution's text to one side and pondered the meaning of the constitution's 'supporting principles and rules', that is, the 'unwritten' constitution.[34] It chose, as Neil MacCormick would say, to 'dig down to the level of constitutional theory'.[35] The Court identified four fundamental organising principles of the unwritten Canadian constitution as particularly relevant to the question of Quebec's secession, namely, federalism, democracy, the rule of law/constitutionalism, and respect for minorities. Only after locating the principles in Canadian constitutional history and expounding their meanings and relationships to each other in abstract terms, did the Court finally confront Question 1 directly. The result is a highly contextualised interpretation of the constitution that reconciles competing constitutional principles and values in such a way as to meet the allegation of Quebec separatists that the textually based answer denying the legality of unilateral secession is contrary to the principle of democracy.

In legal terms, the Court's target was an easy one: the idea that the principle of democracy gives the representatives of a majority of voters in one component of a constitutional order the right to upset unilaterally the entire constitutional arrangement without regard to established legal mechanisms reflects a simplistic view of democracy in which written constitutions would be (as Sir Kenneth Roberts-Wray argued in this context) nothing but scraps of paper.[36] Politically, however, the assertion that *Canadian law* is an illegitimate fetter upon *Québécois democracy* is powerful for some Quebec nationalists. The Court therefore emphasised that, indeed, Canadian federalism allows majorities in each province considerable power to protect and promote their own 'societies' and 'distinct culture[s]'.[37] However, it asserted that democracy is, in the end, much 'richer' than mere majority rule.[38] Democracy also implies respect for individual and minority rights and social justice and equality, and for the democratic expression of popular will to be *legitimate*, it must be made through institutions and methods recognised as *lawful*.[39] The rule of law, by securing a 'stable, predictable and ordered society' in which power is grounded in law, is therefore inseparable from democracy.[40] Moreover, where the constitution is 'entrenched' (as in Canada), the principle of constitutionalism serves to re-define the level of democratic will needed to alter fundamental legal rights or arrangements. Thus, under the Canadian constitution the 'political representatives of the people' are empowered to commit provinces to constitutional rules that protect individual and minority rights from interference by 'simple majority', and these constitutional rules do not frustrate majority rule but define the 'enhanced majority' (or amendment formula) required to be met before such rights may be adversely affected or before the allocation of power between federal and provincial levels is altered.[41]

The Court's analysis on these points represents a classic account of what it means to have an *entrenched* constitution, and how constitutionalism, the rule of law and democracy are complementary, not conflicting, constitutional norms.

34 *Quebec Secession Reference*, n 2 above, para 32.
35 N MacCormick, 'Jurisprudence and the Constitution' (1983) 36 CLP 13, 20 (as quoted by Lord Steyn in *R* v *Home Secretary, ex p Pierson* [1998] AC 539, 584).
36 Sir K. Roberts-Wray, *Commonwealth and Colonial Law* (London: Stevens & Sons, 1966) 293.
37 *Quebec Secession Reference*, n 2 above, paras 58 and 59. Provinces have exclusive legislative jurisdiction over, *inter alia*, 'Property and Civil Rights' and 'Generally all Matters of a merely local or private Nature in the Province'. Constitution Act 1867, ss 92(13) and 92(16)
38 *Quebec Secession Reference*, *ibid* paras 74–76.
39 *Ibid* paras 63–64, 67.
40 *Ibid* paras 70–72.
41 *Ibid* paras 74, 75 and 76.

© The Modern Law Review Limited 1999

However, notwithstanding the legal merit of the Court's analysis, the argument reconciling democracy with the rule of law and constitutionalism may lack moral force for some Quebec nationalists for the simple reason that, in their view, Quebec's 'political representatives' did *not* commit the province to the amendment formulae that now define the 'enhanced majority' needed to change the constitution.[42] The Court, perhaps in recognition of this fact, sought to give further assurances that legality and legitimacy do not diverge in Canadian constitutional law. The solution adopted was to put the unwritten rules of the constitution to work: democracy, it said, implies the need to build majorities through negotiation with and accommodation of 'dissenting voices', and therefore if the *text* of the constitution gives a participant in confederation the right to initiate constitutional amendments then the democratic *context* of the constitution suggests that other parties to confederation ought to negotiate to accommodate this 'dissenting voice', and that this duty to negotiate is a *legal* one.[43] This newly discovered legal duty to negotiate constitutional amendments is clearly significant for Quebec's secession.

The Court's discussion of the fundamental 'unwritten' principles of the Canadian constitution represents a concise and accessible account of some of the basic legal and political principles that underlie many modern liberal democratic constitutional orders. Central to this account is the rejection of any hierarchy of unwritten constitutional norms: the concepts of federalism, democracy, the rule of law and minority rights cannot be defined in isolation and do not compete with one another for supremacy but rather are understood as interwoven and mutually reinforcing norms that share the task of providing political and moral coherence to the whole constitutional and legal superstructure. One purpose of the Court's excursion into constitutional theory was its desire to justify the textual argument about unilateral secession in broader terms of political and constitutional morality: when the Court finally turned to answer Question 1, its conclusion – that secession represents a radical change to the constitution that requires its amendment, and the constitution's amendment formulae preclude unilateral changes even if supported by a majority of one province's voters – appeared perfectly consonant with the legal, political, and moral values that ought to inform the meaning of constitutional democracy in a federal, multinational, multicultural state.

The second, and more important, purpose for the Court's discussion of general principle was to prepare the ground for the 'legal framework' for Quebec's secession. Although it had answered Question 1, the Court proceeded to supplement its answer with conclusions about how Quebec might *lawfully* secede from Canada. The legal framework for secession it constructed may be summarised as follows: if a 'clear majority' of the people of Quebec voted in a referendum on a 'clear question' in favour of secession, federal and other provincial governments could not remain 'indifferent'; rather, 'the federalism principle, in conjunction with the democratic principle' dictates that they would be subject to a *legal* duty to negotiate with Quebec in order to amend the constitution accordingly, and, moreover, all parties to these negotiations would be subject to the *legal* duty to conduct themselves in accordance with the key principles of the unwritten constitution, namely, federalism, democracy, the rule of law and minority rights.[44] Because the unwritten principles share equal status, no single principle could 'trump' the others.[45] Thus, Quebec could not rely upon democracy to insist upon a

42 See text at notes 24–27 above.
43 *Quebec Secession Reference*, n 2 above, paras 68 and 69.
44 *ibid* paras 87–93.
45 *ibid* para 91.

© The Modern Law Review Limited 1999

legal entitlement to secede; its legal right would extend only to *bona fide* negotiations.[46] Conversely, other parties could not rely upon federalism, the rule of law or minority rights to refuse to negotiate secession: as long as Quebec itself 'respects the rights of others' they would be under a legal duty to negotiate with Quebec.[47] To be lawful, secession requires *bona fide* negotiation and agreement aimed at settling difficult questions of, *inter alia*, division of national debt and assets, borders, rights of minorities, including Aboriginal peoples, and future relations between Canada and a sovereign Quebec.[48]

The Court was somewhat unclear about who would have a legal duty (or right) to negotiate secession. It clearly included federal and provincial governments, but it referred generally to 'participants in' and 'parties to' confederation and acknowledged that negotiations would include '*other participants*' than federal and provincial governments.[49] Aboriginal nations participated in building confederation,[50] and, indeed, an Aboriginal right to participate in, or at least be consulted about, secession negotiations is implied by section 35.1 of the Constitution Act 1982, which commits federal and provincial governments to holding a 'constitutional conference' with Aboriginal peoples prior to amending constitutional provisions relating to their rights.[51] Whichever parties are represented directly or indirectly at the negotiations, lawful secession requires constitutional amendment and therefore ratification by entities that may or may not have been directly involved in negotiations; however, the Court refused to indicate which of the two amendment formulae would govern or whether the *unwritten* rules of the constitution require the consent of other parties to confederation than the federal and provincial governments.

Also, the Court was ambiguous about the referendum vote that would precipitate the duty to negotiate secession. What is a 'clear majority' of Quebec's people? Is it a simple majority (50 per cent plus one), or something more?[52] What is a 'clear question' on secession? Would either the 1980 or 1995 questions have qualified?[53] The Court simply stated that the democratic will in favour of secession should be 'free of ambiguity'.[54] Within days of the release of the ruling, political actors began

46 *ibid.*

47 *ibid* para 92.

48 *ibid* para 96.

49 *ibid* paras 88 and 92 (emphasis added).

50 Royal Commission on Aboriginal Peoples, *Partners in Confederation: Aboriginal Peoples, Self-Government, and the Constitution* (Ottawa: Canada Communications Group, 1993) 22.

51 Elsewhere, the Court says secession could not be achieved lawfully without taking 'aboriginal interests ... into account' (*Quebec Secession Reference*, n 2 above, para 139) In general, Royal Commission on Aboriginal Peoples, *Canada's Fiduciary Obligation to Aboriginal Peoples in the Context of Accession to Sovereignty by Quebec* Volume 2 *Domestic Dimensions* (by Reneé Dupuis and Kent McNeil) (Ottawa: Minister of Supply and Services, 1995) 63–65.

52 The PQ government says a 'clear majority' means a simple majority. About 16 per cent of Quebec's population is non-francophone; this group is almost unanimous in its opposition to secession, and so Quebec Deputy Prime Minister Bernard Landry has argued that unless a 'clear majority' means simple majority 'our compatriots, brothers and sisters from the cultural communities', ie non-francophone communities, would have a 'right of veto' 'on our national project' and that 'can't be done' (Tu Thanh Ha and Rhéal Séguin, 'PQ deputy argues 50% plus 1 is enough', *Globe and Mail*, 1 September 1998, A1) The federal government argues that the Court contemplated some higher level of support before the legal duty to enter into secession negotiations arises (E. Greenspon and A. McIlroy, 'Bouchard and PM square off', *Globe and Mail*, 22 August 1998, A1 and A3).

53 On the 1980 and 1995 questions, see n 6 and n 23 above. In a letter to the Prime Minister of Quebec, Lucien Bouchard, the federal Minister of Intergovernmental Affairs, Stephane Dion, stated that, in the federal government's opinion, the 1980 and 1995 referendum questions would not meet the Court's requirement of a 'clear question' on secession (excerpts of the letter are reprinted in *The Vancouver Sun*, 26 August 1998, A11).

54 *Quebec Secession Reference*, n 2 above, para 87.

© The Modern Law Review Limited 1999

disputing the meaning of these phrases.[55] The federal government's view that 'clear majority' represents a higher threshold than 50 per cent plus one is perhaps supported by the Court's language: the Court stated that an 'enhanced majority', not a 'simple majority', is required to amend the constitution,[56] but if it had intended to say that a 'simple majority' could trigger the negotiations required to achieve the 'enhanced majority' needed to amend the constitution, it presumably would not have used another term, 'clear majority', to describe that triggering event.

These and other points of ambiguity might have been clarified in subsequent litigation, but the Court was insistent that such litigation would be inappropriate. The determination of the point at which the legal duty to negotiate secession arises, and the determination of whether parties perform their legal duty to negotiate consistently with the principles of federalism, democracy, the rule of law and minority rights, are both matters over which the Court stated it will not exercise its supervisory jurisdiction: these are non-justiciable matters upon which only political actors would have the 'information and expertise' to make judgments.[57]

Question 2: unilateral secession at international law

The PQ government's position before the *Secession Reference* was that Quebec's secession was not a matter for Canadian law at all, but rather was one of international law. There are, however, serious obstacles to establishing Quebec's *legal right* to secede at international law. Although separatists emphasise the status of *les Québécois* as a 'people' with a right in international law to self-determination,[58] even expert opinions commissioned by the PQ government deny that this right extends to unilateral secession except in extreme cases of colonial or oppressive conditions.[59] Furthermore, by pressing its case on the grounds of the status of *les Québécois* as a 'people', questions are inevitably raised about why all people within the *province* of Quebec should be regarded as belonging to this Quebec *people*, and why Aboriginal *peoples* within Quebec should be denied the same right of self-determination. Emphasis is therefore placed by separatists on a second argument: that international law does not specifically preclude unilateral secession by a component of an existing state, and once secession is established *de facto* it will be acknowledged by the international community in a *de jure* sense under the principle of 'effectiveness'.[60] Using this argument, separatists are able to say that, under international law, Quebec may unilaterally secede with its present borders intact.[61]

---

55 n 52 and n 53 above.

56 *Quebec Secession Reference*, n 2 above, paras 75 and 76.

57 *ibid* para 100.

58 eg D. Turp, 'Quebec's Democratic Right to Self-Determination' in J. McCallum (ed), *Tangled Web: Legal Aspects of Deconfederation* (Toronto: C.D Howe Institute, 1992) 107–115.

59 Cassese, n 11 above, 248–254, N. Finkelstein, G. Vegh, C. Joly, 'Does Québec have a Right to Secede at International Law?' (1995) 74 Can Bar Rev 225; Suzanne Lalonde, Comment (1997) 76 Can Bar Rev 258.

60 On the principle of effectiveness, *Madzimbamuto* v *Lardner-Burke*, n 5 above; T. Honoré, 'Allegiance and the Usurper' [1967] Camb LJ 214; T. Honoré, 'Reflections on Revolutions' (1967) 2 Ir Jur (NS) 268; J. Eekelaar, 'Rhodesia: The Abdication of Constitutionalism' (1969) 32 MLR 34; R.W.M. Dias, 'Legal Politics: Norms behind the *Grundnorm*' [1968] Camb LJ 232.

61 J. Woehrling, 'Les aspects juridiques de la redéfinition du statut politique et constitutionnel du Québec' in Commission on the Political and Constitutional Future of Quebec (Bélanger-Campeau Commission), *Eléments d'analyse institutionnelle, juridique et démolinguistique pertinents à la révision du statut politique et constitutionnel du Québec* (Background papers), vol. 2 (Quebec, 1991) and John McCallum, 'The View from Bélanger-Campeau: Legal Aspects of Quebec Secession' in *Tangled Web*, n 58 above, 73–85; Woehrling, 'Les aspects juridique d'une éventuelle sécession de Québec', n 7 above, 314–328.

 © The Modern Law Review Limited 1999

The Court concluded that while international law recognises the right of 'peoples' to self-determination,[62] this right does not authorise impairing the 'territorial integrity or political unity of sovereign and independent States conducting themselves in compliance with the principle of equal rights and self-determination of peoples'.[63] The right is one of '*internal* self-determination', and it gains an external dimension (including, perhaps, a right to unilateral secession) only where colonial or oppressive conditions exist.[64] Thus, the Court concluded that although 'much of the Quebec population' may constitute a people, it did not need to determine whether that 'people' included all or part of the provincial population, nor did it have to consider the status of Aboriginal peoples, for whatever the scope of the *Québécois* people, it is neither colonial nor oppressed and therefore has no international legal right to unilateral secession.[65]

As for the argument that *de facto* unilateral secession becomes *de jure*, the Court was willing to concede that a 'successful revolution begets its own legality',[66] but then concluded that any such *de jure* recognition 'does not relate back to the date of secession to serve retroactively as a source of a "legal" right to secede in the first place'.[67] The Court also observed that international recognition may be contingent upon adherence by the seceding unit to existing international and domestic legal norms.[68] Indeed, the Court insisted that this fact represents one of the principal ways in which the non-justiciable *Canadian* legal duties relating to secession would be enforced.[69]

Analysis

The ambiguous aspects of the legal framework for secession noted above will likely provoke considerable academic commentary in Canada, but the purpose of the following analysis is not to address those matters. Instead, it will be suggested that the *Quebec Secession Reference* rests upon basic legal, constitutional and jurisprudential assumptions the critical evaluation of which may be important to public lawyers in the United Kingdom and other common law and/or multinational jurisdictions. The issue of Scottish independence provides a particularly useful comparison in the UK context, for in Scotland as in Quebec there exists a long-standing nationalist and separatist movement that is organised politically and that intends to utilise a local, national legislature as a means to further its secessionist aspirations.[70] With respect to the independence of Scotland from the United Kingdom, it can be assumed that, like Quebec, there is no right to make a UDI under either domestic or international law. Thus, were the separatist Scottish

62 *Quebec Secession Reference*, n 2 above, paras 113–116.
63 *ibid* para 120 (quoting UN General Assembly, Declaration on the Occasion of the Fiftieth Anniversary of the United Nations, *GA Res.* 50/6, 9 November 1995).
64 *ibid* paras 126, 131 and 134.
65 *ibid* paras 125 and 135.
66 *ibid* para 144 (quoting S.A. de Smith, 'Constitutional Lawyers in Revolutionary Situations' (1968) 7 West Ontario L Rev 93, 96).
67 *ibid* para 142.
68 *ibid* para 143.
69 *ibid* para 103
70 *Royal Commission on the Constitution, 1969–1973, Report*, Cmnd 5460 (1973), Volume 1, 105–107 and ch 12; Richard J. Finlay, *Independent and Free: Scottish Politics and the Origins of the Scottish National Party, 1918–1945* (Edinburgh: John Donald Publishers Ltd, 1994); Finlay, *A Partnership for Good? Scottish Politics and the Union Since 1880* (Edinburgh: John Donald Publishers Ltd, 1997); Scottish National Party, *Independence in Europe: Make It Happen Now!* (Edinburgh: SNP, 1992).

© The Modern Law Review Limited 1999

National Party (SNP) to win a majority of seats in the new Scottish Parliament and then proceed, without prior agreement with the United Kingdom, to pass an Act declaring Scotland to be independent, the Act would clearly be (from the perspective of UK law) *ultra vires* and void: under the Scotland Act 1998, the UK Parliament remains sovereign over Scotland (section 28(7)) and the Scottish Parliament has no legislative power to alter the Act (section 29(2)(c) and Schedule 4, Part 1, paragraph 4(1)) or to enact legislation on reserved matters, (section 29(2)(b)) including, *inter alia*, the 'the Union of the Kingdoms of Scotland and England' and the UK Parliament (Schedule 5, Part I, paragraph 1(b) and (c)), international relations (Schedule 5, Part I, paragraph 7(1)), or nationality (Schedule 5, Part II, section B6), and thus it would have no power to make a UDI. Parallel arguments could be made about unilateral declarations of independence by a Plaid Cymru-controlled Welsh National Assembly or a republican-dominated Northern Ireland Assembly; the powers of these entities under the Government of Wales Act 1998 and the Northern Ireland Act 1998 respectively exclude the right to make such declarations (indeed, as discussed below, in the case of Northern Ireland that exclusion is express). As with Quebec's independence, the independence of Scotland, Wales or Northern Ireland, to be lawful, would require a 'constitutional amendment', which in the UK context means an Act of the UK Parliament acknowledging the relevant national unit to be a sovereign state. Of course, the process leading to the enactment of such legislation would involve consideration of emotional and complex political matters: just as the Supreme Court of Canada concluded that the decision to pursue Quebec's secession had to be made through the 'political process',[71] so the 1973 Report of the Royal Commission on the Constitution in the United Kingdom concluded that, in relation to Scotland and Wales, the 'paramount reason for demanding or rejecting independence must always be political'.[72] The significance of the *Quebec Secession Reference* is its suggestion that these difficult political considerations need not – indeed ought not – be made in a legal vacuum. Examination of three aspects of the *Quebec Secession Reference* – namely, the importance of unwritten or common law constitutional rules and principles, the use of the concept of justiciability and the relationship between law and politics, and, finally, the theoretical and constitutional problems associated with legal continuity in a revolutionary situation – may be useful in considering whether the normative force of law informs, or ought to inform, the political process of secession in the United Kingdom.

The idea of an 'unwritten' or common law constitution

A striking feature of the *Quebec Secession Reference* is that it is largely a decision of constitutional *common law*. This is remarkable because the Canadian constitution (unlike the UK constitution) is generally assumed to be a written and entrenched constitution. Section 52(1) of the Constitution Act 1982 states that the 'Constitution of Canada' is the 'supreme law of Canada' and that any law inconsistent with it is 'of no force or effect'. What is the 'Constitution of Canada'? Section 52(2) states that it 'includes' a series of positive constitutional instruments therein listed, the most important of which are the Constitution Acts 1867 and 1982. The use of the word 'includes' suggests, however, that this list is not exhaustive. The preamble to the Constitution Act 1867 states that the constitution

71 n 31 above.

72 *Royal Commission on the Constitution*, n 70 above, ch 12, para 497.

© The Modern Law Review Limited 1999

of Canada is 'similar in Principle to that of the United Kingdom' and, as the Court has ruled elsewhere, the Canadian constitution therefore includes 'constitutional principles of a customary nature' derived from the UK constitution.[73] Thus, the many constitutional conventions governing the exercise of the royal prerogative and the relationship between ministers of the Crown and Parliament are part of Canada's constitution. However, the Court also takes the Diceyan view that conventions are not legally enforceable,[74] so while they may form part of the constitution in a broad sense, they cannot be part of the 'Constitution of Canada' that forms the 'supreme law' that prevails over conflicting law; indeed, as the Court states, conventions are not really laws at all.[75] The *Quebec Secession Reference* confirms, however, that there is a branch of the constitution that, like conventions, is customary or unwritten, but which, unlike conventions, is *law*.

Where does this unwritten constitutional law come from? In the *Quebec Secession Reference*, the legal duty to negotiate secession identified by the Court may be said to be a specific example of the more general legal duty to negotiate constitutional amendments, which in turn is a corollary of the right, found in the text of the constitution, to initiate constitutional amendments.[76] But while this textual connection to the legal duty exists, it is a tenuous one, and the Court therefore emphasised the *unwritten* legal foundation for the legal duty to negotiate secession and the legal duty to respect unwritten constitutional principles.[77] These unwritten principles and rules of the constitution do not exist in isolation from the text of the constitution; on the contrary, their principal purpose, said the Court, is to fill gaps left by the text so that the constitution as a whole provides a comprehensive legal framework for government.[78] Indeed, the Court said that the unwritten rules 'emerge' from 'an understanding of the constitutional text itself, the historical context, and previous judicial interpretations of constitutional meaning'.[79] Thus, the unwritten rules do not depend upon any particular positive textual provision but they arise out of a loose juxtaposition of text, historical or customary context and previous case law; interpreted together, these sources may produce sufficient foundation for a *judicially* generated legal rule of substance. Although perhaps not common law in the pure sense, these unwritten constitutional rules may be called *common law constitutional* rules.

The idea of an 'unwritten constitution' securing common law constitutional rights has been used with increased frequency by judges in the United Kingdom.[80]

---

73 *Att Gen Quebec* v *Blaikie et al* [1981] 1 SCR 312, 319–320.

74 A.V. Dicey, *An Introduction to the Study of the Law of the Constitution* (London: MacMillan & Co Ltd, 10th ed, 1960) ch 14; *Patriation Reference*, n 26 above; *Quebec Secession Reference*, n 2 above, para 98.

75 *Quebec Secession Reference*, n 2 above, para 98

76 *ibid* para 69 ('The *Constitution Act 1982* .. by conferring a right to initiate constitutional change ... imposes a corresponding duty' to negotiate amendments).

77 *ibid* para 69 ('This duty is *inherent in the democratic principle* ..' (emphasis added)); and para 88 ('The *federalism principle*, in conjunction with the *democratic principle, dictates* that a clear repudiation of the existing constitutional order and the clear expression of the desire to pursue secession by the population of a province would give rise to a *reciprocal obligation* .. to negotiate .. ' (emphasis added))

78 *ibid* para 32.

79 *ibid.*

80 eg *R* v *Secretary of State for the Home Department, ex p Leech (No 2)* [1994] QB 198; *R* v *Minister of Defence, ex p Smith* [1996] QB 517; *R* v *Lord Chancellor, ex p Witham* [1997] 2 All ER 779; *R* v *Home Secretary, ex p Pierson* [1998] AC 539. In general, T.R.S. Allen, 'Constitutional Rights and Common Law' (1991) 11 OJLS 453; S. Sedley, 'The Sound of Silence: Constitutional Law Without a Constitution' (1994) 110 LQR 270; M. Hunt, *Using Human Rights Law in English Courts* (Oxford. Hart Publishing, 1997) ch 5.

© The Modern Law Review Limited 1999

Laws J, one of the more recent architects of this common law constitution, concedes that the common law theory of constitutional rights in the United Kingdom is inherently ambiguous because there is no 'sovereign text' or 'written constitution' expressly stating that citizens actually have such rights; were there such an instrument there would be no 'conceptual difficulty' in this respect, though difficult questions about the 'true interpretation' of the text would exist.[81] The *Quebec Secession Reference* suggests, however, that in times of extraordinary constitutional and political upheaval even a detailed instrument of positive constitutional law may prove insufficient, and that judges may be forced to fall back upon reasoned analyses of legal principles that are ultimately unanchored in specific positive legal provisions, principles that are instead embedded deep in the customs and traditions that inform our sense of what a legal system in a constitutional democracy ought to be. Indeed, although the Court insisted upon the 'primacy' of the written constitution,[82] much of its ruling seems premised upon a brand of anti-positivism according to which written and entrenched constitutions are not 'sovereign' at all. The implicit message of the *Quebec Secession Reference*, then, is that while a written and entrenched constitution may represent a particularly important and expressive manifestation of legal principle, its jurisprudential foundations are, in the end, a body of customary legal principles the source and content of which are far more subtle and implicit.[83] The case confirms that these underlying constitutional principles are no less *legal* than the written ones that they support, but that, as political allegiance to positive sources of law begins to unravel and the idea of revolution is mooted seriously, adherence to the customary or unwritten constitution will become essential if peace and basic order in society is to be maintained. In other words, when the condition of a legal system is 'pathological', only universal legal principles, not specific written ones from the system itself, will secure the ends of the rule of law.

In the United Kingdom, the process of discovering these unwritten constitutional principles may be more straightforward – there is no need to remove a layer of written constitutional text in order to expose them. Indeed, efforts by Canadian and British judges to identify constitutional rules and principles of a common law or customary nature are analogous, even though one system has a written constitution and the other does not. However, there is a critical difference between the two systems as to what judges can do with these principles once they have found them. Common law constitutional rights in the United Kingdom inform and restrict the exercise of delegated and prerogative executive power as well as the meaning of statutes; but such rights remain subordinate to the express words of Parliament.[84] In the *Quebec Secession Reference*, the Court held that while some rules of the

81 *Witham*, *ibid* 783. For the extra-judicial opinions of Sir John Laws on human rights and the common law, see Laws, 'Is the High Court the Guardian of Fundamental Constitutional Rights?' [1993] PL 59.

82 *Quebec Secession Reference*, n 2 above, para 53.

83 In addressing the international legal principle of effectiveness, the Court quoted with approval the statement by H.W.R. Wade, 'The Basis of Legal Sovereignty' [1955] Camb LJ 172, 196, that 'sovereignty is a political fact for which no purely legal authority can be constituted ...' (*Quebec Secession Reference*, n 2 above, para 142). But between the political facts that support the Canadian constitution and the written text of the constitution is a layer of unwritten constitutional law holding up the text. Says the Court: 'Our Constitution is primarily a written one ... Behind the written word is an historical lineage stretching back through the ages, which aids in the consideration of the underlying constitutional principles. These principles inform and sustain the constitutional text: they are the vital unstated assumptions upon which the text is based' (*ibid* para 49); cf *Reference re Manitoba Language Rights* [1985] 1 SCR 721, 752 ('in the process of Constitutional adjudication, the Court may have regard to *unwritten* postulates which form the *very foundation of the Constitution of Canada*' (emphasis added)).

84 eg *Witham* and *Pierson*, n 80 above.

© The Modern Law Review Limited 1999

unwritten Canadian constitution are non-justiciable conventions, others may give rise to 'substantive legal obligations' of a 'specific and precise nature'.[85] Said the Court: 'The [unwritten] principles are not merely descriptive, but are also invested with a *powerful normative force*, and are binding upon both courts and governments'.[86] If by 'government' the Court meant merely the executive, then the unwritten constitution would have the same legal status as the common law constitution in the United Kingdom; however, the Court tended to use 'government' in a broad sense to include both executive and legislative branches.[87] Indeed, unless it intended that the unwritten constitution prevail over federal and provincial Parliaments, the carefully constructed 'legal framework' for Quebec's secession would crumble: the legal duty to negotiate secession could be circumvented by enactment of a regular federal statute prohibiting secession. In Canada, parliamentary sovereignty has always been limited by constitutional provisions dividing legislative power between components of the federation, and more recently it has been limited by provisions protecting individual and minority rights, but the *Quebec Secession Reference* confirms that parliamentary sovereignty is also limited by certain *common law* constitutional principles related to but distinct from entrenched constitutional texts. On a narrow reading of the Court's ruling it can be argued that the Court has not adopted the view that the common law *simpliciter* prevails over Canadian Parliaments. Rather it can be said that the superiority of any given common law rule over Parliament depends upon its classification as an unwritten rule of the *constitution*; as such, the rule becomes part of the 'Constitution of Canada' and, therefore, a law 'superior' to inconsistent Acts of Parliament pursuant to section 52(2) of the Constitution Act 1982. However, the Court itself did not expressly rely upon this argument, and it may therefore be argued that, in Canada, judges are now prepared to assert common law or unwritten constitutional rules over Parliament not because the constitutional *text* may be interpreted as allowing it, but because in their view certain fundamental legal principles are inherently superior to parliamentary will.

Without a written constitutional text, could courts in the United Kingdom addressing the independence of Scotland, Wales or Northern Ireland articulate a legal framework for secession binding on government and Parliament, or even just government? There is no shortage of statutory precedents for the grant of independence to former colonies and national components of the United Kingdom – the Irish Free State (Constitution) Act 1922 and the Ireland Act 1949 provide a clear example of the independence of (part of) a national unit which was an integral part of the United Kingdom. However, what is novel about the *Quebec Secession Reference* is the imposition of a legal duty to *negotiate* secession if a *majority of the electorate* opts for independence in a *referendum* held by the *government of the relevant unit*. A similar rule – recently enlarged – is arguably already part of UK statute law in relation to Northern Ireland: section 1(1) of the Northern Ireland Act 1998 provides (like section 1 of the Northern Ireland Constitution Act 1973, which was based upon section 1(2) of the Ireland Act 1949) that Northern Ireland shall not cease to be part of the United Kingdom without the consent of 'a majority of the people of Northern Ireland voting in a poll', and it adds in section 1(2) that 'if the wish expressed by a majority in such a poll is that Northern Ireland should cease to be part of the United Kingdom and form part of a united Ireland' the United Kingdom government shall lay before Parliament such

85 *Quebec Secession Reference*, n 2 above, para 54.
86 *Ibid*
87 eg *ibid* paras 72 and 74.

© The Modern Law Review Limited 1999

proposals to give effect to that wish as are agreed upon with the government of Ireland. Unlike Canada, then, the United Kingdom has certain express provisions in positive law addressing the difficult question of the break-up of its constitutional order; at least in relation to Northern Ireland, resort to common law rules might be unnecessary. There are, however, some important differences between the Canadian common law constitutional rule and the United Kingdom's statutory rule. The Canadian rule acknowledges the power of the government of the national unit to decide if and when to hold a referendum on secession, and the rule then requires negotiation and agreement between Canada and the government of that unit, whereas the UK rule empowers the Secretary of State, a minister of the *United Kingdom* government, to call a referendum when *he or she* thinks a majority may wish to secede from the United Kingdom, if at least seven years has elapsed since the last such poll (paragraphs 2 and 3 of Schedule 1 of the Northern Ireland Act 1998), and then the rule contemplates that the government of the United Kingdom will negotiate and reach agreement on secession not with the seceding unit but a *third party* to which the seceding unit is expected to adhere (Ireland). Under the UK rule, the local government of the seceding unit seems to have been cut out of the process. This feature of the UK rule cannot, however, be seen as an attempt by the central government to control unilaterally any future secession process; on the contrary, the Act simply incorporates, verbatim, draft legislation agreed upon by representatives of Northern Ireland's unionist and nationalist parties and the governments of the UK and Ireland in the Belfast Agreement signed on Good Friday, 1998.[88] Thus, in relation to Northern Ireland a judicially imposed framework for secession is unnecessary: unlike Canada and Quebec, the relevant parties were able to agree to their own secession (and reunification) framework.

No explicit statutory provisions on secession exist for Scotland or Wales. The question, however, is whether the secession of these countries is governed by a *common law* rule similar to the one developed in Canada. The referenda that would trigger such a common law rule would likely need some statutory support to be legal. It is arguable that both the Scottish Parliament and the National Assembly for Wales will have the power to hold referenda on secession (whether they will choose to exercise this power is, of course, a separate matter). Under sections 28 and 29 and Schedules 4 and 5 of the Scotland Act 1998, the Scottish Parliament has a general legislative jurisdiction limited only by certain express restrictions and reservations, none of which seem to prevent it from consulting the electorate on the question of secession. The powers of the Welsh Assembly are far more restricted: under sections 21 and 22 and Schedule 2 of the Government of Wales Act 1998 it has only those executive powers transferred by Order in Council or conferred by the Act. The power to hold a referendum to consult the electorate on the question of secession will therefore depend largely upon the terms of the Orders in Council by which executive powers are transferred. The Act itself provides only tenuous support for the power. Section 33 empowers the Assembly to make representations about 'any matter affecting Wales', which arguably includes a representation to the UK government to negotiate secession. However, such a representation, to be credible, would have to be preceded by a referendum vote supporting the idea of secession. Section 36 empowers the Assembly to hold 'a poll' for the purpose of ascertaining the views of the electorate about 'whether or how' any of the Assembly's functions should be exercised, but it expressly excludes the functions

88 *Belfast Agreement* Cm 3883 (10 April 1998), 'Constitutional Issues', Art 1 and Annex A ('Draft Clauses/Schedules for Incorporation in British Legislation') See, Brigid Hadfield, 'The Belfast Agreement, Sovereignty and the State of the Union' [1998] PL 599.

© The Modern Law Review Limited 1999

under section 33. However, section 40 empowers the Assembly to 'do anything' which is calculated 'to facilitate, or is conducive or incidental to' the exercise of any of its functions. It be might be argued that a referendum on secession would be lawful under section 40 as facilitating or as incidental to a representation on secession made under section 33. Alternatively, it might be argued that some other function than that found in section 33 could be linked with section 36 to allow a secession referendum. For example, section 32 empowers the Assembly to do 'anything it considers appropriate to support', *inter alia*, the 'Welsh language' and 'arts, crafts, sport or other cultural or recreational activities'. One method of protecting the Welsh language and culture is to seek Welsh independence; thus, a referendum on secession under section 36 could (arguably) be held as a means of determining how the section 32 powers should be exercised. The difficulty with finding a statutory basis for the right to hold a referendum on secession is, of course, that the whole idea of secession may be regarded as inconsistent with the purpose of the Act, which is to create a democratically elected body for the purpose only of performing *executive* functions on behalf of the Crown (see section 1(3)): the decision to hold a referendum on secession might be considered *ultra vires* in so far as it conflicts with this policy of the Act.[89] The Scotland Act 1998 may escape this restrictive interpretation, since it confers broad *legislative* powers which are (apparently) to be exercised on behalf of the Scottish electorate; hence, a consultative referendum – even on secession – would not conflict with the policy of the Act so long as its purpose is to assist the Scottish Parliament in determining the democratic will of the electorate.

Clearly, some doubts may exist about whether the Scotland Act and the Government of Wales Act permit referenda on secession. It is in this respect that the common law, or unwritten, constitution in the UK becomes critically important, for that common law constitution ought to form the context within which these statutes are interpreted. It is at least arguable that judges could identify a set of unwritten constitutional principles similar to those found in the *Quebec Secession Reference* to construct a legal framework for Scottish or Welsh independence: a principle of respect for national existence could be extrapolated from statute, history and past case law and then combined with the principle of democracy to provide (following the reasoning in the *Quebec* case) a 'right' in favour of a national component to hold its own referendum on independence and to insist upon negotiations on secession with the central government if a clear majority of voters in that component supported independence. Again, the argument may be easier to make in relation to Scotland, where the Acts of Union 1706/7 may be said to reflect an underlying respect for a distinct Scottish national and legal existence, than Wales, where the Acts of Union 1536 and 1543 achieved a more thorough form of legal integration with England; however, it is at least arguable that there exists sufficient historical-legal material from which an unwritten principle of national existence can be extrapolated for both countries.[90] This principle, combined with the principle of democracy, could then be taken to inform the meaning of the Scotland Act and the Government of Wales Act so that, for example, any doubts about the power to hold referenda on secession might be resolved in favour of the power's existence. This power to hold secession referenda would, in light of the status of the Scottish Parliament and Welsh Assembly as

89 *Padfield* v *Minister of Agriculture, Fisheries and Food* [1968] AC 997.

90 For an overview of such legal-historical material, see, for eg, *Royal Commission on the Constitution*, n 70 above. On the effect of the Acts of Union 1536 and 1543 on Welsh national identity, see J. Gwynfor Jones, *Early Modern Wales, c. 1525–1640* (London: St. Martin's Press, 1994) 212–213.

© The Modern Law Review Limited 1999

creatures of statute, have to be seen as statutory, but its existence could be judicially recognised by the Judicial Committee of the Privy Council in exercise of its new jurisdiction to determine the scope of devolved power in Scotland and Wales. In the case of Scotland, the question might come before the Privy Council on a reference by the Lord Advocate, Advocate General or Attorney General of a Bill, prior to its enactment, providing for the referendum to determine if it falls within the legislative competence of the Scottish Parliament (section 33 of the Scotland Act); or, in the case of either Scotland or Wales, the question might reach the Privy Council in the course of litigation (for example, an application for judicial review) as a 'devolution issue', either on appeal from a reference of the issue to a lower court or on a reference directly from the House of Lords (Schedule 6 of the Scotland Act; Schedule 8 of the Government of Wales Act).

Of course, in the *Quebec Secession* case the significant part of the Court's legal framework for secession was not the right to hold a secession referendum (which was not questioned) but the duties governing the conduct of governments *after* a separatist referendum victory. Again, from statutes, history and past case law may be inferred common law principles of respect for national identities and minorities in the UK which, when combined with the principle of democracy, form the basis of a legal duty on the part of the UK government to negotiate secession in response to certain events and according to certain principles. This duty would be regarded as purely common law in source and content, and might be judicially recognised by the regular courts in (for example) a typical judicial review application. There are, however, obstacles to be cleared before its judicial articulation is plausible.

Perhaps the most serious objection to such a common law framework for secession in the UK (aside from the fact that it could not bind Parliament)[91] is that the democratic principle – which was so critical to the development of the legal duty to negotiate secession in the *Quebec* case – has traditionally operated as a *constitutional convention* qualifying parliamentary sovereignty in relation to the independence of colonies and national units from the British empire and the United Kingdom,[92] and conventions are not supposed to crystallise into common law rules.[93] The Canadian Supreme Court operates in what is now a very different legal culture from that of UK courts, and so was able to develop, with reference to a written constitutional text, certain 'unwritten' constitutional rules relating to secession. But would UK courts venture this far into the political realm by developing such common law rules, even if binding only upon the government and not Parliament? Is it conceivable that the High Court would judicially review (for example) the refusal of the Prime Minister or Secretary of State for Scotland to enter into negotiations with the First Minister of Scotland on the issue of Scotland's independence in the aftermath of a separatist referendum victory? These are, as the *Quebec Secession Reference* suggests, two distinct questions; an affirmative answer to the former question need not lead to an affirmative answer to the latter question. Before considering these issues further, the use of the concept of justiciability in the *Quebec Secession Reference* must be considered.

91 Unless, of course, the doctrine of parliamentary sovereignty were reinterpreted along the lines suggested by T.B. Smith, 'The Union of 1707 as Fundamental Law' [1957] PL 99, so as to recognise a law relating to the Scotland-England union superior to the UK Parliament.

92 eg Sir K. Wheare, *The Statute of Westminster and Dominion Status* (Oxford: Oxford University Press, 4th ed, 1949) chs 1 and 2; G. Marshall, *Constitutional Conventions: The Rules and Forms of Political Accountability* (Oxford: Clarendon Press, 1984) chs 10, 11 and 12, Roberts-Wray, n 36 above, 260–301.

93 *Patriation Reference*, n 26 above.

© The Modern Law Review Limited 1999

The concept of justiciability and the relationship between law and politics

In identifying the unwritten rules of the constitution, the Court in the *Quebec Secession Reference* emphasised constitutional history and, in particular, the process of confederation in the mid-1860s leading to the enactment of the BNA Act 1867. This process, it concluded, confirmed the democratic foundations of the Canadian constitution: although the BNA Act was an imperial statute, it was drafted by local elected politicians and enacted at their request rather than imposed by 'Imperial *fiat*'.[94] Of course, this evidence of Canada's democratic roots was simply evidence of the then emerging *convention* that the imperial Parliament would not legislate for Canada except upon its request. Thus, the *Quebec Secession Reference* is an example of judicial use of non-legal conventions to formulate legal rules: the *conventional* rule relating to democracy in Canada's constitutional history formed an important part of the Court's 'understanding' of the constitutional context from which it then derived *legal* rules governing secession. This use of convention to develop law illustrates the changing relationship between law and politics: whereas formerly courts were content to avoid intrusion into high politics by leaving the most fundamental rules of the constitution (on, for example, the dissolution of Parliament, the appointment of Prime Ministers and the royal assent of Bills) to the political realm, the Court in the *Quebec Secession Reference* inserted the normative force of law into the heart of political matters relating to the very existence of the country. On closer scrutiny, however, it may be thought that the Court's insertion of law into politics is more apparent than real.

Asserting 'judicial restraint',[95] the Court concluded that its legal framework for secession is judicially unenforceable. In addressing the issue of justiciability, the Court restated the traditional division in constitutional law between law enforced by courts and conventions enforced through political sanction, but it then introduced a further division: even in relation to the '*law* of the Constitution' judicial enforcement may be inconsistent with the Court's proper role in the constitutional order.[96] The Court then concluded that the legal obligations identified in relation to Quebec's secession fall within this category: the judiciary would have 'no supervisory role over the political aspects of constitutional negotiations' for questions about whether the legal duty to negotiate secession had arisen and whether parties conduct the negotiations according to their legal duties to respect the key principles of the constitution involve matters of political judgment that 'def[y] legal analysis'.[97] The Court insisted that these legal duties are 'binding' and that their breach would have 'serious legal repercussions', for where there are legal rights there are 'remedies', but also maintained that the appropriate recourse would be 'the political process rather than the courts'.[98]

If the Court has *no* supervisory jurisdiction over the legal rights and duties associated with secession, then is it the case that these rights and duties are *legal* at all? There are two pragmatic reasons why the Court insisted on describing the rules relating to secession as 'law'. First, although the duties relating to secession could have been, like conventions, described as rules of constitutional morality, perhaps they could not really have qualified as conventions because they did not arise from

94 *Quebec Secession Reference*, n 2 above, para 35.
95 *Ibid* para 99.
96 *Ibid* para 98 (emphasis in text).
97 *Ibid* para 101.
98 *Ibid* para 102.

© The Modern Law Review Limited 1999

a practice or custom that political actors regard as normative; they are clearly judicially created and imposed rules not customary norms.[99] Second, the Court seemed very concerned to meet the separatist argument from democracy by denying any gap between legality and (political) legitimacy, and to describe the framework for secession as anything but 'legal' would have left such a gap – the law would have failed to accommodate Quebec's democratic will to secede. In essence, the Court sought to extend *law* into this gap without having the *courts* follow. There were practical reasons for this approach. Like the UK constitution, the Canadian constitution contains many gaps between law and legitimacy (eg the Crown's prerogative power to refuse assent to Bills or to dissolve Parliament), but, traditionally, these gaps have been addressed by convention (eg royal prerogatives are exercised upon the advice of ministers responsible to a democratically elected House of Commons); but with the increased juridification of constitutional politics in Canada over the past fifteen years,[100] the use of extra-legal norms to address gaps between law and legitimacy is no longer politically persuasive. To help diffuse a tense political situation in which the Court's legitimacy was at stake in Quebec, the Court chose to reconcile law and democratic legitimacy without the aid of extra-legal norms. The question, however, is whether the Court succeeded in doing this in a substantive manner, or whether it simply resorted to semantics. In other words, is there any jurisprudential justification for its description of the rules governing secession as 'law'?

Positivists in the tradition of Bentham, Austin and Kelsen would deny to the Court's rules on secession the status of law because they are not backed by coercive judicial sanction, and even Hart might be sceptical about their status as law given the absence of a 'rule of adjudication'.[101] Underlying the Court's approach, then, is a rejection of positivism and the adoption of a concept of law that does not depend on judicial enforcement or adjudication. The jurisprudential coherence of this approach requires that it is possible to count as 'law' a norm that, like a convention or other customary norm, is enforced through political or social sanction, but which is not merely a convention or other customary norm. In defending the Court's approach it should be emphasised that the law often recognises legal duties or rights that are (in Fuller's term) 'polycentric' and therefore not susceptible to adjudication in a judicial setting.[102] Indeed de Smith and Brazier list a number of statutory and written constitutional provisions that, because they state broad political aspirations, are regarded as non-justiciable – including, significantly, the rule examined above that Northern Ireland shall not cease to be part of the United Kingdom without the consent of a majority of the people voting in a poll.[103] But these examples are all found in positive legal instruments and (it may be said) are 'legal' instead of 'political' for that reason alone. However, in developing the legal framework for secession, the Court did not purport to act as legislature and did not regard the normative force of the legal framework for secession *as law* as deriving from the mere fact that it decided to

99 eg Jennings' definition of conventions Sir I. Jennings, *The Law and the Constitution* (London: University of London Press, 3rd ed, 1943) 131.

100 eg R. Penner, 'The Canadian Experience with the Charter of Rights: Are there Lessons for the United Kingdom?' [1996] PL 104, 116–123.

101 In general, N. MacCormick, 'Coercion and Law' in *Legal Right and Social Democracy. Essays in Legal and Political Philosophy* (Oxford: Clarendon Press, 1982) 236; Hart, *Concept of Law*, n 1 above, 93–95.

102 L. Fuller, 'The Forms and Limits of Adjudication' (1978) 92 Harv L Rev 358

103 S. de Smith and R. Brazier, *Constitutional and Administrative Law* (Harmondsworth: Penguin, 8th ed, 1998) 32.

 © The Modern Law Review Limited 1999

call it 'legal' rather than conventional, political or moral. Rather, the Court presumed that the normativity of these rules *as law* arose because they were derived from a reasoned interpretation of the Canadian constitutional order, including its history, customs, traditions and organising legal principles, and because they fit within this normative matrix of rules and principles in such a way as to best promote the structure as a morally and politically coherent legal system.[104] Norms thus identified may be labelled as *law* and the possibility of coercive sanction may be considered merely as 'an addendum, an additional security' – that is, practically important but not always essential to a norm's status as law.[105] While future litigation about the application of the legal norms governing secession may be precluded, that these norms are *legal* is significant: the character of constitutional discourse in Canada has been fundamentally altered from assertions (and threats) of a political or moral nature to include in addition, reasoned interpretations and arguments about *law*. It would seem that the normative force of these legal arguments is felt by political actors on both sides in very different ways from the normative force of political and moral arguments, notwithstanding (or perhaps because of) the absence of judicial sanction. Thus, the PQ government's traditional claim that Canada would have to enter into negotiations after a separatist referendum victory, which previously was a claim of a political or moral nature that the federal government denied, can now be seen to have the support of law, and, significantly, both sides have accepted the implications of this aspect of the Court's ruling; the nature of the debate has now shifted to arguments about the legal test set by the Court to determine when the negotiations must be held and how they might be conducted.[106] Paradoxically, the very prospect of future judicial intervention into this politically-charged discourse might well have undermined the legitimacy of the Court and destroyed the normative force that law without the threat of judicial sanction is capable of exerting on the relevant political actors. Finally, it should be mentioned that even if the duties are not directly justiciable, courts may take into account the legality of Canada's or Quebec's conduct in deciding whether to give allegiance to the old or the new regime in the event of a UDI by Quebec – indeed, perhaps this is what the Court meant when it said that violations of the duties would have 'legal repercussions'.

If courts in the United Kingdom were reluctant about intervening in sensitive political decisions about the independence of Scotland or Wales, they could take a similar approach: they could make a declaration that the common law, interpreted in light of statute, history and past case law, requires secession negotiations after a victory by separatists in a referendum held by a future Scottish or Welsh executive or Parliament/Assembly, and that there is a common law duty to conduct those negotiations according to certain principles, but then decline supervisory jurisdiction over the interpretation and application of those legal duties.[107] To be effective, a judicial declaration of non-justiciable common law duties relating to secession would have to be unambiguous and, preferably, made by the House of Lords (as a duty imposed on the UK it could not be considered a 'devolution issue' within the jurisdiction of the Privy Council). Political actors might well respond

104 eg R. Dworkin, *Law's Empire* (Cambridge: Harvard University Press, 1986) 225–312.

105 MacCormick, 'Coercion and Law', n 101 above, 240; J. Finnis, *Natural Law and Natural Rights* (Oxford: Clarendon Press, 1980) 276–281

106 Rhéal Séguin, 'Federalist cause poisoned by ruling, Bouchard says', *Globe and Mail*, 22 August 1998, A3 and references at n 52 and n 53 above

107 cf *obiter dicta* recognising the existence of a non-justiciable 'fundamental' law arising from the English-Scottish union binding on government and Parliament alike: *MacCormick* v *Lord Advocate* 1953 SC 396, 412; *Gibson* v *Lord Advocate* 1975 SC 134, 137; *Murray* v *Rogers* 1992 SLT 221, 227.

© The Modern Law Review Limited 1999

differently to such declarations than they would to judicial declarations of convention or constitutional morality. Arguably, the normative effect of such declarations would be felt on both sides. For example, once such a judicial declaration was made it would be far more difficult for the SNP to claim (as it has in the past) that it could insist upon secession negotiations with the United Kingdom government as soon as it won a majority of seats in a parliamentary election.[108] A 'clear' separatist victory in a referendum on a 'clear question' would be necessary. Conversely, it would be far more difficult for the United Kingdom government to claim an exclusive power to determine if and when a referendum would be held – as mentioned, the Privy Council could, interpreting the Scotland Act and the Government of Wales Act in light of the common law rules on secession, conclude that it was competent for the Scottish and Welsh executives and/or Parliament/Assembly to make their own provision for referenda on independence. In short, difficult questions about the mechanics of independence that previously were subject to claims and counterclaims of a purely political character could, after a statement by the courts of a common law framework for Scottish or Welsh independence, be subjected to a more principled, rational and disciplined sort of political discourse informed by a common understanding of basic legal requirements. At the very least, some of the more obviously extreme claims on either side would be confirmed as beyond certain parameters of law set for this complex and emotional political debate.

'Secession' as a legal concept and problems of legal continuity, revolution and autochthony

The Court's opinion in the *Quebec Secession Reference* is premised upon the idea that secession can be achieved *lawfully* – that is, that a legal system may break into two legal systems without a break in legal continuity, or a *revolution*. Under this approach, the legal and constitutional system of an independent Quebec would (apparently) derive its validity from the Canadian legal and constitutional order from which it had separated. Indeed, the Court cited Canada's own independence from the United Kingdom as an example of how a state can become 'fully independent' from another state consistently with the concept of 'legal continuity' and the demands of the 'rule of law'.[109] The PQ government's position is, however, quite different. As seen, in 1995 it was prepared to make a UDI and thus establish Quebec's independence through revolution, but its preferred course was to secede through negotiations with Canada. Although such negotiations might have led to amendments to the Canadian constitution to reflect Quebec's departure, separatists in Quebec would not have considered these amendments as the legal fountain from which their new sovereign legal order would spring. On the contrary, the PQ government stated that *after* it negotiated secession with Canada it would draft a new constitution for Quebec the validity of which would derive from a confirmatory referendum; this expression of popular will would be 'l'acte de naissance du Québec souverain'.[110] The 'acte de naissance' would have been, in other words, the assertion by the new sovereign Quebec state of autochthony, or (as Sir Kenneth

108 eg SNP, *Independence in Europe*, n 70 above, 3; SNP, *Manifesto–General Election 1997*(Edinburgh: SNP, 1997) 9 Among the 'policy intentions' that the SNP will publish prior to the elections for the first Scottish Parliament is the promise of a referendum on independence within the first four years of a SNP administration (see http.//www snp org uk/liberty/policy.html#5).

109 *Quebec Secession Reference*, n 2 above, paras 46 and 47.

110 Parti Québécois, *Programme du Parti québécois* (Montréal: Parti Québécois, 1991) 18 (quoted by R.A. Young, *The Secession of Quebec and the Future of Canada* (Montreal and Kingston: McGill-Queen's University Press, 1995) 117)

 © The Modern Law Review Limited 1999

Wheare said) a 'home-grown' constitutional root.[111] Indeed, the proposed assertion of autochthony by Quebec resembles the SNP programme for Scotland. After negotiating independence with the United Kingdom government, the SNP would hold a 'national referendum' to allow the electorate 'to approve a new written Constitution for Scotland'; the SNP makes no mention of the necessity for an Act of the UK Parliament granting independence.[112] Although Plaid Cymru has traditionally avoided the language of separation and independence,[113] its programme for Wales is similar. It seeks a Welsh Parliament, similar to that which Scotland now has, which could then adopt a constitution for an independent Wales; this constitution, which would be based upon the inherent sovereignty of the Welsh people, would come into force upon ratification by the electorate in a referendum, whereupon Wales would take its place on the international stage with (it is argued) full membership in the European Union, the Commonwealth and the United Nations.[114]

It is often assumed that an autochthonous constitution can only be achieved by revolutionary act, so that to obtain a local root for a new legal system connections with the old system must be severed in a manner not recognised as lawful by the old system.[115] Kelsen went so far as to argue that if the independence of a new state is gained through the laws of a parent state then the new state's laws can be traced back through a chain of validity to the parent state's laws and therefore the two systems are still in fact one; therefore only an unlawful break with the parent state can secure a separate legal order.[116] The implications of this assumption are, however, troubling: a national unit desirous of an autochthonous constitution may seek that objective by refusing to acknowledge and participate in the procedures for independence that the existing system regards as lawful, perhaps making a UDI and thereby threatening peace, order and stability. A better jurisprudential explanation for secession would be to argue, as does John Finnis, that a law of a parent state providing for the independence of a new state remains in force and relevant to the continued legality of the legal system it creates, but that it is not in force in a forward-looking sense and its existence does not diminish the fact that a true and complete severance of systems has taken place.[117] Thus, 'one can accede to the lawyerlike demand for a root of title without imperilling the independence and autochthony of a new state'.[118] Says Finnis, the Kelsenian view that an autochthonous legal system cannot emerge where independence was achieved through the laws of a parent state is unsound because the separate identity of a legal system results not from the absence of technical legal links with another legal system of which it was formerly a part but from the independent social and political identity of the community it serves.[119]

111 Sir K. Wheare, *The Constitutional Structure of the Commonwealth* (Oxford Clarendon Press, 1960) ch 4.

112 SNP, *Independence in Europe*, n 70 above, 3; SNP, *Manifesto–General Election 1997*, n 108 above, 9.

113 Thomas Christiansen, *Plaid Cymru in the 1990s Dilemmas and Ambiguities of Welsh Regional Nationalism* (Florence European University Institute, Paper No 95/3, 1995) 8, 34–36.

114 Plaid Cymru, *General Election Manifesto* (1997); see also, Plaid Cymru's position on the 1997 devolution referendum campaign at http·www.plaidcymru.org/policy/yesDW.htm.

115 Wheare, n 110 above; Roberts-Wray, n 36 above, 289–301.

116 H. Kelsen, *General Theory of Law and State* (Cambridge: Harvard University Press, Eng tr, 1946) 115–117, J. Raz, *The Concept of a Legal System: An Introduction to the Theory of Legal System* (Oxford. Clarendon Press, 1970) 102–108; Raz, 'The Identity of Legal Systems' in Raz, *The Authority of Law: Essays on Law and Morality* (Oxford: Clarendon Press, 1979) 99.

117 J. Finnis, 'Revolutions and Continuity of Law' in A.W.B. Simpson (ed), *Oxford Essays in Jurisprudence (Second Series)* (Oxford: Clarendon Press, 1973) 64. See also Peter Oliver, 'The 1982 Patriation of the Canadian Constitution· Reflections on Continuity and Change' (1994) 28 *Revue Juridique Thémis* 875.

118 *ibid.*

119 *ibid* 67–69. Also, Raz, *Concept of a Legal System*, n 116 above, 188 and Raz, 'Identity of Legal Systems', n 116 above, 99–100; G. Maher, 'The Identity of the Scottish Legal System' [1977] Jur Rev 21, 36–37.

© The Modern Law Review Limited 1999

Indeed, unless the possibility of *non-revolutionary autochthony* is accepted, one is driven to look for revolutions where none really occurred in order to buttress claims to autochthony. Thus, a distinction may be drawn between autochthony by *real* revolution – for example, a UDI made with no intention of seeking prior agreement with the parent state – and autochthony by *theoretical* or *paper* revolution – for example, *after* independence is achieved by some method recognised as lawful by the parent state, officials in the new state reinterpret the grounds of independence as being another method not recognised as lawful by the parent state. These various theories of autochthony are illustrated by interpretations of Canada's independence from the United Kingdom. In its ruling in the *Quebec Secession Reference* the Court concluded that Canada achieved *de facto* sovereignty, with the United Kingdom's consent, sometime prior to the 1926 Balfour Declaration, but that this sovereignty was not complete in *legal* terms until the enactment by the UK Parliament of the Statute of Westminster 1931 and the Canada Act 1982.[120] However, comments by the Court in two other parts of the ruling suggest that, in its view, the validity of the Canadian constitution does not *now* derive from these British statutes. At one point it quoted with approval H.W.R. Wade's assertion that 'sovereignty is a political fact for which no purely legal authority can be constituted',[121] and at another point it stated that the Canadian constitution 'is the expression of the sovereignty of the people of Canada'.[122]

It may be said, then, that the Court asserted an *autochthonous* basis for the Canadian constitution – its ultimate root is not Westminster but the local political fact of popular sovereignty. But was it asserting revolutionary or non-revolutionary autochthony? In the article by Wade to which the Court referred (albeit in a different context) it is argued that the independence of dominions like Canada from the United Kingdom was achieved by a 'revolution' hidden under 'legal dress': local judges in such dominions subtly shifted their allegiance from an imperial to a local constitution that denied the most basic tenet of UK law, parliamentary sovereignty, such that they would not uphold as lawful any future UK legislation reasserting authority over the dominion even though UK courts would.[123] In light of this weighty academic authority in favour of the revolution thesis, it is perhaps surprising that the Court chose to invoke the example of Canada's independence from the United Kingdom in an opinion in which the notion of secession as a legal concept was pivotal. However, there is a very good explanation for why the Court's use of Canada's independence was not inconsistent with its objectives, an explanation that reveals much about judicial priorities in times of political upheaval and uncertainty.

As a matter of historical fact, the relevant parties involved in procuring the various independence statutes for Canada *thought* that they were conducting themselves according to legal rules, and thus the process was stable and orderly and consonant with the demands of the rule of law.[124] If Canada's independence involved a revolution at all it was a *paper* one only. Since paper revolutions result from interpretations of law that discount the credibility of legal appearances it is fair to require that before they are accepted as good interpretations of law they

120 *Quebec Secession Reference*, n 2 above, paras 46 and 47.

121 *ibid* para 142 (quoting Wade, n 83 above, 196).

122 *ibid* para 85.

123 Wade, n 83 above, 190–91. See also, B Slattery, 'The Independence of Canada' (1983) 5 Sup Ct L Rev 369, 390–93, 398–99.

124 E. McWhinney, *Canada and the Constitution, 1979–1982* (Toronto: University of Toronto Press, 1982) 66–69; Hurley, n 33 above, chs 2 and 3.

© The Modern Law Review Limited 1999

overcome a presumption that legal appearances match legal substance. In other words, one may ask why a revolutionary interpretation of Canadian autochthony should be accepted if a non-revolutionary explanation is possible. And, such a non-revolutionary explanation is possible. If Parliament is indeed sovereign and can make any law except one that binds its successors then it can grant independence to a colony, and, although it can always repeal that grant and reassert its authority, unless and until that possibility becomes reality the grant must be regarded, in law as well as in fact, as effective and complete – Parliament must be taken at its word and the legality of full independence acknowledged.[125] If the new state subsequently exercises its full sovereignty to define its constitution as locally rooted, or autochthonous, then it is only exercising powers lawfully granted and cannot be said to have behaved in a revolutionary manner. It is only if Parliament repeals the grant and reasserts authority over the former colony that the need for a 'revolution' arises; however, in that case, 'revolution' would not be an appropriate term to describe the likely resistance of authorities in the former colony to the assertion of British authority: the assertion by the UK Parliament of authority over Canada or Australia, for example, would not now differ in kind from its assertion of sovereignty over France, and refusal by French authorities to accept British legal authority over their country could hardly be described as a 'revolution'.[126] Although the Court in the *Quebec Secession Reference* did not attempt develop such an argument, it is clear that (notwithstanding its reference to Wade's article), it concluded that Canada's autochthony was non-revolutionary.

The theoretical debate surrounding Canada's independence from the UK derives from the troublesome implications of the doctrine of parliamentary sovereignty, and, although relevant to the question of Scottish, Welsh or Northern Irish independence from the UK, these problems are not directly relevant to Quebec's secession from Canada; in Canada the constitution, not federal and/or provincial Parliaments, is supreme. However, analysis of competing interpretations of Canada's independence illustrates the important distinction between revolutionary and non-revolutionary autochthony, and the distinction between real and paper revolutions. For a new state, revolutionary autochthony may represent a politically compelling form of constitutionalism, for it implies a clear repudiation of the validity of an existing legal order that may well be reviled; however, where the existing constitution contains within it a 'legal framework' for secession that respects both democracy and the rule of law, this political desire for revolutionary autochthony can be satisfied by a paper revolution that, because it is only a matter of legal construction arising *after* independence, need not prevent an orderly, peaceful dismemberment of the existing state. Indeed, Wade's revolution in 'legal dress' achieves the objectives of peace, stability and order inherent in the ideal of the rule of law and is, from the perspective of the parent state, not really a revolution (either paper or real) at all. In 1995 the PQ government in Quebec contemplated making a UDI that would have established autochthony by real revolution, thus risking peace and order, but its preferred route to autochthony was, like the SNP's programme for Scotland, a negotiated agreement followed by a localised constitution-building exercise leading to the birth of a new sovereign state. The significance of the *Quebec Secession Reference* is that by providing a legal framework for secession, the Court has removed any justification that might

125 *British Coal Co* v *R* [1935] AC 500, 597; *Murray* v *Parkes* [1942] 2 KB 123; *Att Gen Ontario* v *Att Gen Canada* [1947] AC 127, *Ibralebbe* v *R* [1964] AC 900, 924; *Madzimabamuto* v *Lardner-Burke* [1969] 1 AC 645, 722; *Manuel* v *R* [1982] 3 All ER 786, 792–794.

126 *Manuel, ibid.*

© The Modern Law Review Limited 1999

exist in political or moral terms for asserting autochthony by real revolution. Of course, even if Quebec were to obtain independence through this legal framework, authorities in Quebec might still wish to reinterpret independence as revolutionary. However, the above analysis of the *Quebec Secession Reference* confirms two important points in this respect. First, such a paper revolution would not be necessary in order for Quebec's constitution to be autochthonous – indeed, 'l'acte de naissance du Québec souverain' described in the PQ programme is quite consistent with non-revolutionary autochthony. Second, the possibility of such revolutionary interpretations in Quebec would not undermine the legality of Quebec's secession in Canadian law and would not, therefore, concern the Supreme Court of Canada – the judicial responsibility for upholding the rule of law would, by this point, already have been met as best as the situation permitted.

The conclusion that autochthony and the rule of law are not irreconcilable is particularly important for the United Kingdom. Returning to the question posed above – should UK courts articulate a legal framework for secession? – it can be seen that by inserting the normative force of law into such a highly-politicised question an opportunity is created for democracy, or legitimacy, to operate in tandem with legality, and in the event that democratic forces make independence an inevitability, the chances of autochthony being established in a non-revolutionary or paper revolutionary manner are improved. Of course, in the legal and political culture of the United Kingdom today, it is likely that parliamentary sovereignty coupled with constitutional conventions designed to promote democracy would suffice.[127] It is premature to consider the need for judicial intervention in relation to either Scotland or Wales, and the incorporation of a negotiated secession framework into the Northern Ireland Act 1998 suggests that the UK government is willing to agree with other interested parties about the basic rules governing the secession of national units when political events make such rules necessary and, in any event, before judicial intervention is either possible or required. It is only in the unlikely event that the UK government and a Scottish or Welsh separatist executive cannot agree on a framework for secession that the articulation of such a framework by the courts may be necessary to ensure that the emergence of a newly independent and autochthonous state, if that must be the end result, will be achieved without *real* revolution and the accompanying threat to order and stability. Perhaps the *Quebec Secession Reference* is significant for the UK not so much as a model for courts to follow, but as a reminder to governments that their conduct in relation to secessionist claims should be governed by basic unwritten legal rules and principles relating to democracy, the rule of law, and respect for national identities and minority rights, and that unless they negotiate amongst themselves a secession framework when necessary a court may just possibly intervene to impose one.

127 cf J. McFadden and W. Bain, 'Strategies for the Future: A Lasting Parliament for Scotland?' in T. St. J.N. Bates (ed), *Devolution to Scotland: The Legal Aspects* (Edinburgh. T. & T. Clark, 1997) 16, who argue that a mere statement of policy from the government adopting for Scotland a rule on secession similar to that enacted in legislation for Northern Ireland might well suffice.

© The Modern Law Review Limited 1999

LEGISLATION

The Crime and Disorder Act 1998: Child and Community 'Safety'

Christine Piper*

The provisions of the Crime and Disorder Act 1998 ('the Act') relating to children and young people came as no surprise to those with academic and professional interests in youth justice and crime prevention, such provisions being based on measures previously outlined in the White Paper, *No More Excuses – A New Approach to Tackling Youth Crime in England and Wales,*[1] and in the Audit Commission's Report, *Misspent Youth, Young People and Crime.*[2] However, the scope and significance of those provisions may have escaped more general notice, partly because the title does not suggest a focus on children and young people and also because debate, in Parliament and the media, in the weeks immediately before the Act received the Royal Assent on 31 July 1998, was dominated by controversy around an amendment, which was not enacted, to lower the age of consent for homosexual sex from 18 to 16 years.[3]

The Act as a whole provides new court orders relating to the prevention and punishment of crime and disorder, amends the criminal law and revises aspects of the youth and adult criminal justice systems. However, the Act's title does not give any indication that so many of the Act's provisions are aimed at minors. Family lawyers, for example, may only now be noting that the Act has provisions which, when implemented,[4] will make new orders available to the Family Proceedings Court. Indeed, one might argue that if the Act had been entitled the 'Children and Young Persons' Act' that would not have been so very misleading, given the focus of such a large proportion of the Act[5] on those under 18 years of age.[6] The many provisions of the Act relating to children and young people have the potential, as had the Children Act 1989, for changing professional philosophies for working with children and families and for influencing how we 'see' children.

* Law Department, Brunel University.

1 Cm 3809 (London: Stationery Office, 1997). This White Paper was itself based on consultation documents issued by the Government when in Opposition and during its first six months in power. See, for example, *Safer communities, safer Britain, Labour's proposals for tough action on crime* (London: Labour Party, 1995); J. Straw and A. Michael *Tackling Youth Crime. Reforming Youth Justice, A consultation paper on an agenda for change* (London: Labour Party, May 1996) and *Community Safety Order, A Consultation Document* (London: Sentencing and Offences Unit, Home Office, September 1997).

2 London: Audit Commission, 1996.

3 This proposed reform was part of a new clause 1 which had been passed in the House of Commons by a majority of 207 on 22 June 1998 but was rejected in the House of Lords on 22 July by a majority of 168 and dropped by the Government in order to ensure the passage of the Act: see R. Swade 'The Crime and Disorder Bill, The Age of Consent Debate' (1998) 148 *Childright* 6.

4 Pilots for child safety orders began in selected areas on 30 September 1998 and will run for 18 months with the aim of national implementation in 2000–2001.

5 As V. Baird points out, 'there are youth-related measures in all five parts of the Act' ('Crime and Disorder Act 1998: introduction and new preventative orders in youth justice' (1998) *Legal Action* October 15, 15).

6 The Act uses the previous statutory definitions of 'child' as under 14 and 'young person' as 14 and under 18 (s 117)

© The Modern Law Review Limited 1999 (MLR 62 3, May). Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

Targeting children and young people

Of obvious importance is the abolition, by section 34, of the rebuttable presumption that a child over 10 years of age is *doli incapax* – a presumption previously applying to 10–13 year olds.[7] Provisions in the Act relating to, for example, racially aggravated offences (in sections 28–33) are, therefore, potentially relevant to children over 10.[8] Furthermore, the Act creates a new order – the anti-social behaviour order – which is available to the courts in relation to anyone aged 10 or over in England and Wales and aged 16 or over in Scotland.[9] Drug treatment and testing orders will be available in both jurisdictions for those aged 16 or over.[10]

There are also provisions relating specifically and exclusively to children and young people. Sections 11–15, dealing with child safety orders and child curfew schemes and notices, relate only to the under 10 year olds; section 16, allowing the removal of truants to designated premises, relates to children and young persons of compulsory school age; parenting orders are triggered when a child or young person is the subject of one of a range of orders or is convicted of an offence.[11] In addition there are provisions relating to the setting up of statutorily based systems to 'process' children and young people who offend: sections 65–66 set up a new system of reprimands and final warnings to replace police cautions, sections 37–42 provide a statutory framework for a youth justice system with services, plans and inter-agency teams, and a national Youth Justice Board is set up by section 41 and schedule 2.[12] Existing provisions for remands and committal of children and young people are amended by sections 97 and 98, the court's ability to take account of the accused's silence at trial is extended down to 10–13 year olds[13] and, for those found guilty of offending, the courts will have additional or re-named sentencing options: a reparation order,[14] an action plan order[15] and a detention and training order,[16] together with new sanctions (under sections 72 and 77) for breach of supervision requirements.

On the other hand, very few sections of the Act relate explicitly and exclusively

7 See, for a discussion, S. Bandali 'Abolition of the Presumption of Doli Incapax and the Criminalisation of Children' (1998) 37(2) *Howard Journal* 114.

8 Those sections of the Act which relate to new sex offender orders (ss 2–3 for England and Wales and s 20 for Scotland) and extended sentences for licence purposes for sex and violence offences (ss 58–60 and 86–88) might also apply to young people – given the extension of culpability in relation to sexual offences; see, for example, the Sex Offenders Act 1997 s 4.

9 By ss 1 and 19, respectively. None of the provisions in the Act relate to Northern Ireland except certain of the provisions abolishing the death penalty for treason and piracy (s 36).

10 ss 61–64 and 89–95, Schedules 4 and 6.

11 ss 8–10 deal with parenting orders. section 8(1) lists the relevant orders as a child safety order, an anti-social behaviour order (created by s 1 of the Act) and a sex offender order (created by s 2 of the Act). In addition a parenting order can be triggered by the child failing to comply with a school attendance order under ss 443–444 of the Education Act 1996 (s 8(1)(d) of the Act) and by a criminal conviction (s 8(1)(c) of the Act).

12 The Act also provides time limits within the prosecution process for the under 18 year olds (s 44) and amends the powers of youth courts (ss 47–48).

13 s 34 of the Act amends s 35 of the Criminal Justice and Public Order Act 1994 by removing the words 'who has attained the age of fourteen years' from s 35(1) and by deleting all of s 35(6).

14 ss 67–68. Alternatively, under s 71, the court can impose a reparation requirement in a supervision order.

15 ss 69–70 and Schedule 5.

16 ss 73–79. For a discussion of these new orders in the context of the current use of the Children and Young Persons Act 1933 s 53(2) and (3), see C. Ball '*R* v *B (Young Offenders: Sentencing Powers)*. Paying due regard to the welfare of the child in criminal proceedings' (1999) 10(4) *Child and Family Law Quarterly* 417.

© The Modern Law Review Limited 1999

to 'grown-ups'. The abolition, by section 51 and Schedule 3, of committal proceedings for indictable only offences relates to 'an adult'; section 99, which inserts a new section into the Criminal Justice Act 1991 and thereby amends powers to release short term prisoners on licence, relates to those 'aged 18 or over' though the new section 34A(5)(a) gives the Home Secretary the power to repeal the words 'aged 18 or over'. Those sections of the Act which are not explicitly aimed at children and young people are nearly all[17] concerned with sexual and violent offenders,[18] offenders dependent on drugs[19] and offenders motivated by race.[20]

In relation to adults, and those younger citizens who are brought within the scope of that term, the Act is, therefore, concerned with what have become, in legislation over the last two decades, 'the usual suspects': those offenders who are perceived as the most 'dangerous' to society at this point in time[21] because they are sexually or physically violent, use or traffic in drugs or threaten the stability of a multi-racial society.[22] In this context, an Act entitled 'Crime and Disorder' which concentrates to the extent this Act does on children and young persons is clearly endorsing those political and social ideas which emphasise the 'danger' of young people's behaviour – the perceived threat from children 'out of control' and the potential threat to society if children are not guided into responsible and law-abiding adulthood.

Images of children

Such fears are, of course, not new:[23] they have justified state intervention in factories and families since the early 19th century with a focus on one or both of those two categories of children labelled by the Victorians as the 'depraved' and the 'deprived'.[24] They have, however, currently been amplified by a particular 'politicisation' of law and order[25] whereby, according to von Hirsch and Ashworth, 'popular resentment' about crime has been exploited by both right and centre left

17 The exceptions to this generalisation include amendments to police powers of seizure (ss 24, 26–27), powers to remove masks (s 25), bail conditions and restrictions (ss 54–56), the abolition of the death penalty for treason and piracy (s 36), sentencing guidance (ss 80–1) and the 'toughening up' of penalties for failure of compliance by football spectators with reporting restrictions under s 16(5) of the Football Spectators' Act 1989 (s 84).

18 See ss 2–4 and 20–22 (sex offender orders) and ss 58–60 and 86–88 (sentences extended for licence purposes for sexual and violent offenders).

19 See ss 61–64 and 82–95 and schedules 4 and 6 (drug treatment and testing orders).

20 See ss 28–33 (racially aggravated offences) and ss 82 and 96 (sentencing increases for racial aggravation) See Maleiha Malik, 409 below.

21 See J. Pratt 'Governing the Dangerous. An Historical Overview of Dangerous Offender Legislation' (1996) 5 *Social and Legal Studies* 21 for an analysis of changing conceptions of 'dangerousness'.

22 See, for example, the Criminal Justice Act 1993 and the Crime (Sentences) Act 1997 for provisions relating to violent, sex and drug offenders and the Criminal Justice Act 1991 for the provision in ss 1 and 2 whereby those convicted of particular sexual and violent offences can be sentenced to 'longer than normal' terms of imprisonment.

23 See, for example, G. Pearson *Hooligan: A History of Respectable Fears* (London: Macmillan, 1983); H Hendrick *Child Welfare, England 1872–1989* (London: Routledge, 1994) especially ch 1; J. Eekelaar, R. Dingwall and T. Murray 'Victims or Threats? Children in Care Proceedings' (1982) *Journal of Social Welfare and Family Law* 68.

24 See, for example, A. Morris and H. Giller *Understanding Juvenile Justice* (London: Croom Helm, 1987) ch 1; C. Piper 'Moral campaigns for children's welfare in the nineteenth century' in M. King (ed) *Moral Agendas for Children's Welfare* (London. Routledge, 1999) ch 3.

25 The politicisation of criminal justice policy as such is not novel: it has, for example, occurred from time to time in relation to juvenile justice from at least the 1960s onwards: see J. Pitts *The Politics of Juvenile Crime* (London: Sage, 1988).

© The Modern Law Review Limited 1999

political parties and governments in successfully 'garnering political support':[26] 'What is different about "law and order" in the 1980s and 1990s is the stridency of the appeals to fear'.[27] An example of such an appeal can be found in Jack Straw's Labour Party report of April 1996 entitled *Tackling disorder, insecurity and crime,* which begins:

> There is a rising tide of disorder which is blighting our streets, parks, town and city centres and neighbourhoods. ... Disorder has been all too often ignored. ... yet it has profound effects on individuals who feel frightened and unsafe: it can help to tip whole areas into decline, economic dislocation and crime; it can undermine the commercial viability of town and city centres.

The fears engendered have increasingly focused on the young in the 1990s, despite unclear statistics about juvenile offending trends[28] and despite the fact that the death of James Bulger in 1993 – which intensified concern and led to calls for toughter penalties for young offenders – was unrepresentative of youth crime.[29] But, as Diduck argues, 'Moral concern [for children] can easily turn to moral condemnation, particularly in a political context in which the rhetoric of morality trips off the tongues of both the new right and the new labour communitarianism'.[30] According to Littlechild one can pinpoint when this occurred: 'It is possible to put a date on the genesis of the new punitive policies in the youth justice system: at the time when the police and the media were on the hunt for the rat boy, bail bandits and ram raiders; and when the Conservative Party was low in the opinion polls'.[31] For the same political reason, there began in 1996 'what seemed to be like a Dutch auction of who could seem to be most tough in relation to young people'.[32]

The responsible child

What has emerged – sometimes in contradiction, as we shall see, to the smaller print of policy documents – is a particular image of the young offender which concentrates on the personal responsibility of the child or young person for his or her offending. *Tackling Youth Crime,* having outlined the proposed reforms which are now in the Crime and Disorder Act 1998, made the new approach clear: 'Changes of this kind ... would represent a move away from preoccupation with legal processes to one of confronting young offenders with their behaviour'.[33] Straw, as Home Secretary, re-iterated this theme in *No More Excuses*:

> An excuse culture has developed within the youth justice system. It excuses itself for its inefficiency, and too often excuses the young offenders before it, implying that they cannot help their behaviour because of their social circumstances. Rarely are they confronted with their behaviour and helped to take more personal responsibility for their actions.[34]

---

26 A. von Hirsch and A Ashworth 'Law and Order' in A von Hirsch and A Ashworth (eds) *Principled Sentencing* (Oxford: Hart Publishing, 2nd ed, 1998) 410, 413–414.

27 *ibid* 411

28 See B. Littlechild 'Young offenders, punitive policy and the rights of children' (1997) 17(4) *Critical Social Policy* 73, 78–79; M. Hill and K. Tisdall *Children and Society* (London: Longman, 1997) 178–180.

29 Venables and Thompson (the two boys convicted in this case) were sentenced under the Children and Young Persons Act 1993 s 53(1) which provides for the equivalent of a life sentence. The sentencing powers available were, therefore, adequate though whether it is adequate to deal with children in the Crown Court is another issue.

30 A. Diduck 'Justice and childhood. reflections on refashioned boundaries' in M. King (1999) n 24 above, ch 8.

31 n 28 above, 77

32 *ibid* 80.

33 Straw and Michael, n 1 above, 12.

34 White Paper, n 1 above, Preface by the Home Secretary.

 © The Modern Law Review Limited 1999

According to this White Paper, 'Children above the age of criminal responsibility are generally mature enough to account for their actions and the law should recognise this'.[35] They are, therefore, deemed mature enough to respond to measures focusing on 'nipping offending in the bud':[36]

> Many young people offend once or twice, often under the influence of their peers. The challenge is to nip that behaviour in the bud and to prevent further criminality ... a range of community sentences must be provided in every area to prevent the first time offender from developing into a repeat offender.[37]

This 'no-excuse' culture and the constitution of children and young people as sufficiently responsible and mature to be held to account has also allowed restorative justice to re-emerge as a legitimate option in relation to offending by minors. Mediation and reparation projects established and researched within the juvenile and criminal justice systems in the 1980s had revealed more difficulties than advantages. By 1992 Davis had concluded, 'It is rare for a set of ideas to catch on as quickly as did the enthusiasm for victim-offender mediation and reparation in the UK in the mid 1980s ... It is now fair to say that, within government circles, mediation and reparation schemes constitute something of a "dead" subject'.[38] The political need since 1992 to be seen to be tackling youth crime, against a backdrop of criticisms of the current youth justice system, has given reparation a new lease of life. It is presented as a new, more positive means of responding to offending by the young and, through its elements of blaming the offender whilst promoting reintegration into the community, appears to cross the punishment/protection divide.

Research undertaken within the social sciences is called in aid of these images and possibilities. *No More Excuses* asserts in its Preface: 'For too long we have assumed that young offenders will grow out of their offending if left to themselves. The research evidence shows this does not happen' and paragraph 1.10 of that document[39] accurately summarises recent Home Office research which shows that young men are now taking much longer to desist from offending and that those traditional indicators of 'settling down', such as marriage, no longer correlate with such desistance (though they are still operative in regard to young women).[40] However, that research also shows that 'offending amongst young people is widespread and ... most offenders commit no more than one or two offences', with three per cent of such offenders counting for about a quarter of all offences,[41] a finding distorted in the White Paper: 'For many young offenders it is true that their first caution – or court appearance – is enough to divert them from crime. But this assumption is wide of the mark when it comes to the hard core of persistent offenders who cause so much crime'.[42] 'Most' has changed to 'many' and the focus of the message is on the 'hard core': measures and language suitable for a minority set the tone for the majority.

35 *ibid* Introduction.

36 *ibid* Preface: 'We are determined to cut waste in the present Youth Justice System as identified by the Audit Commission last year. Instead we will refocus resources and the talents of professionals on nipping offending in the bud, to prevent crime from becoming a way of life for so many young people'.

37 *Safer communities, safer Britain*, n 1 above, 6–7.

38 G. Davis *Making Amends: Mediation and Reparation in Criminal Justice* (London: Routledge, 1992). For schemes operating in the 1980s see, for example, H. Blagg 'Reparation and Justice for Juveniles' (1985) 25 *British Journal of Criminology* 269; M. Wright and B. Galaway (eds) *Mediation and Criminal Justice* (London: Sage, 1989).

39 n 1 above.

40 J. Graham and B. Bowling *Young People and Crime* HORS 145 (London: Home Office, 1995).

41 *ibid* 83 and 84.

42 *No More Excuses*, n 1 above, para 1.9.

© The Modern Law Review Limited 1999

The dominant image of the young offender is, therefore, not of a child but of a (male) youth: sections 8–16 of the Act, dealing with parenting, child safety and curfew orders, all available for the under 10 year olds and two exclusively so, are entitled 'Youth crime and disorder'. As the Bulger case revealed, we find it difficult to conceptualise as 'normal' those children who offend – they become 'evil monsters' at one end of the spectrum and youth at the other[43] and, if von Hirsch and Ashworth are right in their analysis of symbolic criminal justice politics,[44] politicians have led, rather than followed, public opinion in these constructions. But that analysis also warns that there are 'subspecies' of such politics:

> One such subspecies seems to be a certain mixture of nostalgia and notions of 'instilling discipline', represented by measures such as curfews for young teenagers. The children involved are not necessarily characterised as vermin ... the idea, rather, seems to be the instillation of a certain 'benevolent discipline'.[45]

That rather different image of a naughty child was evident in *Tackling Youth Crime*. Having summarised the now familiar message about the need for young offenders 'to be held to account' the following statement is made: 'All this is common sense. It is how people deal with unacceptable behaviour by children in a family setting or at school or on the sports field'.[46] Such homely comments have clear appeal: that they gloss over the known difficulties of substitute child care and control by state agencies is, therefore, a source of some concern.

The vulnerable child

These current dominant images of the minor who offends or engages in 'deviant' behaviour, though by no means novel, are in contrast to other images which have underpinned legislative provisions in the past. For example there has at times been an image of the child whose offending is viewed as an indicator of need – whether that need is in relation to social and economic conditions or personal behaviour difficulties. In the 1970s and 1980s policies of diversion from prosecution and from custodial sentences, developed under the influence of financial constraints and social science research and theory, were based on ideas about the deleterious effect of labelling[47] children as criminals and the propensity for offending to be a stage in the development of children. Such ideas are evident in the White Papers preceding the Social Work (Scotland) Act 1968 and the Children and Young Persons' Act 1969[48] and also in the provisions in the Children Act 1989 to impose duties on Local Authorities, in the context of children 'in need', to encourage children not to offend and to divert them from criminal proceedings.[49]

The dominant image of the young offender is also in contrast to images of children underpinning recent legislation relating to the family. For example, the Family Law Act 1996 relies on an image of a vulnerable child harmed by parental conflict and violence and who needs protection from divorcing or abusive parents

43 See, for example, M King *A Better World for Children* (London: Routledge, 1997), chapter entitled 'The James Bulger trial: good or bad for guilty or innocent children?'.
44 n 26 above, 419.
45 *ibid.*
46 n 1 above, 9.
47 See E. Lemert *Human Deviance, Social Problems and Social Control* (New York: Prentice Hall, 1967).
48 Home Office *The Child, The Family and the Offender* Cmnd 2742 (London: HMSO, 1965); Home Office *Children in Trouble* Cmnd 3601 (London: HMSO, 1968).
49 s 17 and Sched 2 para 7

© The Modern Law Review Limited 1999

by state encouragement of mediation and remedies to remove abusers from the family home. Likewise, in the family justice system, so strong is the image of the dependent, vulnerable child that court welfare officers are often reluctant to 'burden' children with the opportunity to express their wishes and judges are reluctant to give leave to minors to make independent applications.[50] Even when the child's liberty is at stake in an application for her to be placed in secure accommodation under section 25 of the Children Act 1989 the child has no right to be in court because it is deemed harmful for her.[51]

Child and community protection

And yet there are passages in policy documents that would found a different image of the child or young person who offends and would influence less punitive practice in relation to them. One such passage can be found in the Introduction to the recent Consultation Document *Supporting Families*:[52] 'First the interests of the child must be paramount. The Government's interest in family policy is primarily an interest in ensuring that the next generation gets the best possible start in life'. Whilst this may be constituting childhood simply as preparation for adulthood it does concentrate on the child's welfare. However, this is point 7, whereas point 1 gives a different message: 'Families are at the heart of society. ... Our future depends on their success in bringing up children. That is why we are committed to strengthening family life'. The ambiguity as to whether the aim is to promote the child's welfare *per se* or to protect adults is also evident in the Government's apparent acceptance[53] of the research which finds statistical correlations between the onset or persistence of offending and various factors concerned with family life, learning difficulties, unemployment and drug or alcohol use[54] and its enacted 'solutions' which focus on final warnings and a range of court imposed orders.

The Labour Government has, therefore, given mixed messages about the relationship between child and community 'safety'. In Opposition the position appeared clear: 'Ultimately the welfare needs of the individual young offender cannot outweigh the needs of the community to be protected from the adverse consequences of his or her offending behaviour'.[55] In the White Paper this conflict is resolved differently: 'Preventing offending promotes the welfare of the individual young offender and protects the public. ... Preventing offending by young people is a key aim: it is in the best interests of the young person and the public'.[56] So, using this version of best interests which trumps all other child welfare needs, the section on child safety and curfew orders can be headed 'Protecting children under 10'.[57]

Nevertheless, in Standing Committee, the responsible Minister (Mr O'Brien) stated; ' this Bill is concerned with protecting both the child and the community'

50 See chapters by J Roche and C Piper in S. Day Sclater and C. Piper (eds) *Undercurrents of Divorce* (Aldershot: Ashgate, 1999) ch 3 and ch 4, respectively.

51 *Re W (minor) (secure accommodation order: attendance at court)* [1994] 2 FLR 1092.

52 London: Home Office, 1998, 4.

53 See, for example, *No More Excuses*, n1 above, ch 1.

54 See, for example, D. Utting, J. Bright and C. Henricson *Crime and the Family*, Occasional Paper 16, (London: Family Policy Studies Centre, 1993) and M. Rutter, H. Giller and A. Hagell *Antisocial Behaviour by Young People; The Main Messages from a Major Review of the Research* (Knutsford: Social Information Systems Ltd, 1998) for reviews of such research

55 Straw and Michael, n 1 above, 9.

56 n 1 above, paras 2.2. and 2.3.

57 *ibid*, before para 5.2.

© The Modern Law Review Limited 1999

and 'The [child safety] order's purpose will be to facilitate a balance'.[58] Child protection has generally avoided such notions of balance though there are important exceptions. One of these[59] is the Children Act 1989 provision mentioned above which empowers the court to make a secure accommodation order.[60] The criteria for that order include, in section 25(1)(b), 'likely to injure himself or other persons' which creates a potential conflict between the interests of the community and the child and has led to the exclusion of the principle that the welfare of the child should be paramount.[61] Whether the courts will be able to give sufficient weight to the child's welfare in the balancing exercise created by child safety orders is similarly problematic, a point made at length by the Opposition in Standing Committee. The Minister's reply was not wholly reassuring:

> The Hon Member for Hertsmere is correct in saying that the paramount consideration is for the care of the child. The orders will also allow consideration of the impact of the child's behaviour, which could be damaging to its victims and to the wider community. Such behaviour could be damaging to the child, if there was no intervention to stop it. [Mr Clappison asks for clarification.] As the trigger for a care order under the clause would be the child's original behaviour followed by the breach of an order made under the Bill, the court would have to have regard to that, as well as to the care of the child. ... The child's welfare would be neglected by our not stopping such behaviour.[62]

These attempts to conflate child and community safety raise a related issue, that of the separation or otherwise of the young 'criminal' from the young 'victim'. The trend, since the failure to implement much of the Children and Young Persons Act 1969,[63] has increasingly been to 'process' the child who offends through law and procedures quite separate from those relevant to the child in need of services or protection. This has occurred partly as a result of the punitive attitudes already discussed but also as a result of longer standing concerns to minimise state intervention in the lives of both groups of children and to set up appropriate procedures for them. Examples of this increasing separation are the repeal by the Children Act 1989 of the criminal offending condition for a care order[64] and the division of the old juvenile court into separate Family Proceedings and Youth Courts by section 70 of the Criminal Justice Act 1991. The deprived and depraved are now normally responded to differently, there remaining very few pieces of legislation or sites[65] which bridge the distinction between them. Child safety orders seem set to muddy the distinction.

---

58 HC Standing Committee B 7 May 1998 cols 273 and 270 respectively.

59 Another example is part IV of the Family Law Act 1996 which deals with protection of children, spouses and cohabitees from domestic violence: the court, in deciding whether to exercise its discretion to make an occupation order, may have to balance the harm such an order would inflict on the respondent (or his child) against the harm the applicant (or her child) might suffer if an order were not made. See, for example, ss 33(7)(b) and 36(8)(b).

60 s 25, operating in relation to a child 'looked after' by the Local Authority, that is, in care under s 31 or accommodated under s 20 of the Children Act 1989. The scope was extended by the Children (Secure Accommodation) Regulations 1991 regulation 7 to children accommodated by health authorities and trusts and in non-Local Authority residential care and mental nursing homes.

61 *Re M (Secure Accommodation Order)* [1995] 1 FLR 418. See C. Smith and P. Gardner for the difficulties caused by these conflicting aims: 'Secure Accommodation under the Children Act 1989: legislative confusion and social ambivalence' (1996) 18(2) *Journal of Social Welfare and Family Law* 173.

62 Mr Alun Michael, HC Standing Committee B col 287 and 288 12 May 1998.

63 See, for example, R. Harris and D. Webb *Welfare, Power and Juvenile Justice* (London: Tavistock Publications, 1987).

64 s 90 of and Sched 15 to the Children Act 1989 repealed section 7(7) of the Children and Young Persons Act 1969.

65 As mentioned above, the Children Act 1989 Sched 2 deals with the prevention of offending in the context of need, and admission to secure units can be by welfare (s 25 of the Children Act 1989) and criminal processes.

© The Modern Law Review Limited 1999

That is not necessarily something to be criticised. Valid arguments for conflating the categories and not labelling the child who offends as 'an offender' *per se* have been put forward by politicians, juvenile justice practitioners and charitable bodies from time to time over at least 200 years and have, occasionally, formed the dominant view. However, those pursuing such an approach must develop practices which fully acknowledge the dangers such a blurring of distinctions can produce for the child and her family and formulate best practice accordingly. There are grounds for concern that this might not occur in relation to child safety orders.

Fudging distinctions: the child safety order

On the application of a local authority, the family proceedings court will be able to make a child safety order in relation to children under 10 years of age if one or more of the conditions specified in section 11(3) are fulfilled:

(a) that the child has committed an act which, if he had been aged 10 or over, would have constituted an offence;
(b) that a child safety order is necessary for the purpose of preventing the commission by the child of such an act as is mentioned in paragraph (a) above;
(c) that the child has contravened a ban imposed by a curfew notice;[66] and
(d) that the child has acted in a manner that caused or was likely to cause harassment, alarm or distress to one or more persons not of the same household as himself.

These are 'fairly wide grounds'[67] and cover conceptually different scenarios. Conditions (a) and (b) reveal the political imperative to 'deal with' an ever younger category of (potentially) offending children. Children over 10 will be liable to reprimands, warnings and criminal proceedings in respect of their offending; the under 10s may instead be the subject of a child safety order if offending (without the *mens rea* element) is proved (without the child having any right to separate representation) to the satisfaction of the court (on the balance of probabilities). Conditions (c) and (d) appear to be targeted at the child who, in public spaces, is a nuisance or who is deemed to be in such places at unsuitable hours for a child. The fourth condition is identical in its wording to part of section 1 of the Act. That section provides for anti-social behaviour orders for anyone over 10 years of age and includes a definition of acting in an 'antisocial manner': 'that is to say, in a manner that caused or was likely to cause harassment, alarm or distress'.[68] In effect, this means that a child safety order, imposed on this condition, amounts to an anti-social order for the under 10s.

The order is for a maximum period of three months[69] and places the child under the supervision of a 'responsible officer' who can be a member of a local authority youth offending team (established under section 39) or a social worker in the social

66 s 11(3)(c) will not be available for the courts until child curfew schemes have been authorised and become operational

67 Mr Clappison, HC Standing Committee B 7 May 1998 col 274. An amendment to substitute the following for the first three grounds. 'that there are reasonable grounds for believing that the child's behaviour indicates a serious risk of the child offending after he has attained the age of 10 and that making such an order is in the interests of preventing the child so offending' was not pressed to a division

68 s 1(1)(a), replicated in s 11(3)(d).

69 Unless the circumstances are 'exceptional' in which case the maximum is 12 months

© The Modern Law Review Limited 1999

services department of a local authority[70] and 'requires the child to comply with such requirements as are so specified'.[71] Those requirements, set out in section 11(5), are any[72] 'which the court considers desirable in the interests of – (a) securing that the child receives appropriate care, protection and support and is subject to proper control; or (b) preventing any repetition of the kind of behaviour which led to the child safety order being made'.

In contrast to the Family Law Act 1996, section 11(4) does not impose its own welfare test or checklist, and so the welfare of the child will be paramount in proceedings considering applications for these orders.[73] The application must be made by a local authority and the child will probably be supervised by a social work trained officer. It may, therefore, be seen as alarmist to view such orders as heralding an Orwellian state. However, this focus on offending to trigger a civil order – without the child-friendly provisions built into the child protection process or the procedural safeguards built into the criminal process – is worrying. For example, applications for child safety orders will not be 'specified proceedings' so a guardian *ad litem* will not be appointed for the child, nor is the child entitled to separate legal representation (because she is not a party, nor are they criminal proceedings). Similarly, whilst the child safety order is essentially a supervision order (with requirements) it does not have the same safeguards built into its use. A supervision order is available within the family justice system via section 31 of the Children Act 1989 where the 'significant harm' threshold operates,[74] and within the youth justice system under section 7(7)(b) of the Child and Young Persons Act 1969. The problem, as *Childright* has commented, is that 'the extra care, protection and support' which a child safety order might give to a child may only be useful 'so long as they are not branded as criminals and excluded from society'.[75] Child safety orders could label and criminalise children to their detriment and bring more children into youth justice and care systems when less draconian responses might be more appropriate. There is no requirement to consider (as in section 1(5) of the Children Act 1989) whether, even if the conditions have been satisfied, it is better for the child not to make an order and so it is not clear how the magistrates' discretion in section 11(1) ('may make an order') will be exercised.

Perhaps more serious, however, are the implications for the child (and for the coherence of family law) who fails to comply with any requirements included in a child safety order. The order can be varied or, alternatively, it can be discharged and a care order made in its stead under section 31(1)(a) of the Children Act 1989. Section 90(1) of the Children Act 1989 abolished the power to make care orders on an offence condition because, as Masson notes, 'The offence condition ... did not fit in with the notion that a care or supervision order should only be made on proof of substantial harm',[76] and yet the Crime and Disorder Act 1998 explicitly states that the significant harm test to be found in the section 31(2) of the Children Act does not have to be applied.[77] Before the introduction of the Children Act 1989

70 s 11(1)(a) and (8).

71 s 11(1)(b)

72 Except that s 12(3) specifies that requirements which conflict with the parent's religious beliefs and the child's schooling should be avoided.

73 As applications for child safety orders are family proceedings, the welfare test in the Children Act 1989 s 1 operates rather than the test in the Children Act 1933 s 44 ('have regard to the welfare of the child or young person') which operates in criminal proceedings.

74 Though it is fair to say that a s 31 order gives the Local Authority parental responsibility for the child (s 33(3)).

75 The Children's Legal Centre (1998) 143 *Childright* 8, 9.

76 J. Masson *Current Law Statutes Annotated* (London: Sweet & Maxwell, 1989) Children Act 1989, 41–159

77 s 12(7).

© The Modern Law Review Limited 1999

there were at least 12 routes into care[78] and section 12 of the Crime and Disorder Act 1998 does not technically alter the intention of the Children Act 1989 to establish only one route into care but, by specifically excluding the necessary operation of the threshold test, in effect it does create a different route into care.[79]

The potential of the provision for child safety orders is that an instance of actual 'offending' or the risk of such, the failure to 'stay in' between the hours of 9 pm and 6 am as specified by the terms of a curfew notice[80] and instances of anti-social behaviour could all lead to the imposition of conditions on a child and its family which, if not complied with, allow a court to by-pass that very test that was felt necessary in 1989 to reduce the incidence of the most severe form of state intervention in the family. Furthermore, behaviour which falls short of offending (anti-social behaviour by anyone over or under 10, together with 'offending' or the risk of such by those under 10 deemed incapable of offending) can, therefore, potentially lead, depending on the child's age and the original order, to the removal of a child from his or her family to be 'looked after' either by the Local Authority or by HM Prison Service.

The partnership context

This ambiguity as to whether the primary aim is the promotion of the child's welfare or crime prevention is of significance in relation to an important theme in the Act – that of inter-agency partnership. Under section 39 Local Authorities must set up youth offending teams, and probation committees and police and health authorities must cooperate in such provision and in the making of youth justice plans for the area. In practice many such teams already exist, having developed *ad hoc* over the last decade or so and inter-agency partnerships are routine in child protection work.[81] What undermines government assurances that community protection can be properly pursued without a reduced emphasis on child protection and support is that it has been seen as necessary to set up a new and separate partnership. Schedule 2 to the Children Act 1989, together with the Children Act 1989 (Amendment) (Children's Services Planning) Order 1996, requires the local authority to produce and publish children's services plans and such plans should include provision for services to prevent offending behaviour by children and young people. Yet a consultation paper issued in early 1998 by the Department of Health, entitled *Working Together to Safeguard Children: New Government Proposals for Inter-Agency Co-operation*,[82] did not mention youth offending. One could, therefore, argue that new 'dedicated' partnerships are necessary because crime preventative measures have not been given adequate priority and resources

78 P. Bromley and N Lowe *Bromley's Family Law*, (London: Butterworths, 7th ed, 1992) 440 *et seq.*

79 Significantly, seven years before the partial repeal of s 7(7) of the Children and Young Persons Act 1969 (see above) a new sub-section has been inserted into that section to enjoin the courts to use it only in relation to serious offending and only if the child was also 'in need of care or control which he is unlikely to receive unless the court makes a care order' (s 7(7A) inserted by the Criminal Justice Act 1982 s 23).

80 s 14(2); s 15(4) amends s 47 of the Children Act 1989 so that the Local Authority also has a duty to investigate the child's welfare if informed that a curfew notice has been contravened.

81 For government encouragement of partnership in probation practice see, for example, *Partnership in Dealing with Offenders in the Community* (London: Home Office, 1990) and *Supervision and Punishment in the Community* Cm 966 (London: HMSO, 1990). For such encouragement in relation to child protection see, for example, F. Kaganas, M. King and C. Piper (eds) *Legislating for Harmony, Partnership under the Children Act 1989* (London: Jessica Kingsley Publication, 1995)

82 London; Children's Services Branch, Department of Health, February 1998.

© The Modern Law Review Limited 1999

in children's services plans[83] but that could have been remedied. How children at risk of offending or who do offend are protected will depend at least partly on the nature of the partnership between the partnerships.

How the new orders and processes provided by the Crime and Disorder Act 1998 are 'contextualised', resourced and marketed will, therefore, be crucial. If preventative measures are developed, presented and resourced as opportunities and services sufficiently widely available not to stigmatise and alienate then they are greatly to be welcomed. Similarly, if child safety, anti-social behaviour and parenting orders include requirements that provide child and family support rather than negative control they too can open up new avenues for helping childen and their families. Otherwise these measures will simply provide different technical ways to draw children and families into youth justice and child protection systems which often do not improve their lives.

One suspects that, as with the introduction of community service orders by section 14 of the Powers of Criminal Courts Act 1970, the lack of penological and conceptual clarity is not an accident. The Act can be all things to all people. It appears to legitimate much wider possibilities for victim-offender and community-offender reparation and encourages the sort of inter-disciplinery working that could allow the child to be seen as more than simply an offender. Yet these provisions sit alongside what is in effect a tariff-based, rather than child-centred, system of reprimands and warnings and alongside civil orders whose requirements for 'detention' in the home could be increasing the dangers to a child. How youth justice professionals and magistrates approach and use these orders and services will be a major determinant of outcome, given the inherent ambivalence in government policy.

The proclaimed commitment of the Labour Party and Government to a two-pronged strategy in relation to young offenders is well known: 'Both research and common sense indicate that there are links between social conditions and crime ... As well as tackling these underlying causes of offending, immediate action can be taken at a local level to reduce crime'.[84] It would be a pity if the immediate action in the provisions of the Crime and Disorder Act 1998 set the tone for professional and magisterial responses to children who offend. Terms such as risk, harm and safety are now ubiquitous concepts deployed in different policy areas but with different meanings and outcomes in each. There is a danger that the Act will contribute to the dominance of particular meanings whereby children's rights and welfare are subsumed in community safety. If that occurs, then it is doubtful whether the child, in experiencing the potentially restrictive and controlling nature of these new orders, will appreciate that he or she is being 'saved' from a life of depravity.

83 See Social Services Inspectorate/Department of Health *Responding to Youth Crime: Findings from inspections of youth justice services in five local authority social services departments* (London: HMSO, 1994); B. Hearn and R Sinclair *Children's Services Plans, Analysing need· reallocating resources*, (London: Department of Health, 1998).

84 *Safer communities, safer Britain*, n 1 above 8.

© The Modern Law Review Limited 1999

'Racist Crime': Racially Aggravated Offences in the Crime and Disorder Act 1998 Part II

Maleiha Malik*

The introduction of racially aggravated offences in Part II of the Crime and Disorder Act 1998 (the Act) represents a major shift in the state response to violence and harassment of minorities in the United Kingdom (UK).[1] It comes at a time when 'racial violence' is increasingly recognised as a persistent social problem. High profile incidents such as the Stephen Lawrence enquiry have ensured that the debate about the relationship between racial violence, crime and law enforcement agencies in the United Kingdom has taken on greater urgency.[2] Previous governments resisted pressure to introduce a specific criminal offence of racial violence and harassment. In 1994, the Home Secretary, Michael Howard summarised this approach in setting out the Government's objection to the Home Affairs Committee's call to introduce such an offence in its report on Racial Attacks and Harrassment: 'All violent crimes, regardless of motivation, can already be dealt with properly under existing legislation'.[3] In a sharp contrast to this approach, the Labour Party election manifesto stated that 'Britain is a multiracial and multicultural society. All of its members must have the protection of the law. We will create a new offence of racial harassment and a new crime of racially motivated violence to protect ethnic minorities from intimidation.'[4]

The introduction of racially aggravated offences brings the UK in line with the strategy of other members of the European Union (EU) and the United States of America (USA). Some individual member states of the EU have amended and supplemented their existing criminal law provisions in order to respond to the distinct form that racism has taken in these countries. The new Criminal Code in France which came into force on 1 March 1994, amended and supplemented existing provisions on racism. In relation to the offence of desecration of graves, heavier penalties are now incurred when these offences are racially motivated.[5]

* School of Law, King's College London.

I should like to thank Lupin Rahman, Jonathan Glover and the Centre for European Law (KCL) for their assistance.

1 For the most recent figures on the incidence of racial violence and harassment see *'Ethnicity and Victimisation: Findings From the 1996 British Crime Survey*, Home Office Statistical Bulletin Issue 6/98, Research and Statistics Directorate (Government Statistical Service, 3 April 1998).

2 See *The Stephen Lawrence Enquiry: Report of an Enquiry by Sir William Macpherson of Cluny* (London: Home Office, February 1999).

3 The Government Reply to the Third Report From the Home Affairs Committee Session 1993–94 HC 71 (1994), Cm 2684. See HC Deb vol 235 col 32 11 January 1994 for Michael Howard's rejection of the Committee's proposals. There were a number of attempts to introduce legislation. See HC Deb vol 215 col 850 9 December 1992 for the attempt by Mr David Winnick to make racial violence and harassment a specific criminal offence. See also HL Deb vol 556 cols 1646–67 12 July 1994 for proposals by Lord Irving of Lairg to introduce a new offence of racially motivated violence; and *ibid* cols 1682–1699 for Baroness Flather's suggestion for statutory recognition of racial motivation as an aggravating factor in sentencing.

4 *New Labour· because Britain deserves better*, Labour Party Manifesto 1997, (London: Labour Party, 1997) 23.

5 See Article 225–17 of the new Criminal Code 1994. This followed the desecration of Jewish graves in a cemetery in Carpentras in 1990 (*The Times*, 15 August 1990). For a discussion of these changes see European Commission against Racism and Intolerance (ECRI), 'Legal measures to combat racism and intolerance in the member States of the Council of Europe', Strasbourg, 21 March 1996.

© The Modern Law Review Limited 1999 (MLR 62:3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

The German government has introduced legislation to extend criminal liability for neo-Nazi, racist and xenophobic attacks.[6] In addition, solutions are being sought at a regional level as the institutions of the EU, and especially the European Parliament, are increasingly pre-occupied with the problem of 'racism and xenophobia'.[7] The USA has seen an explosion in the number of state legislatures which have introduced hate crimes legislation into their penal codes.[8] They have either introduced a specific criminal offence where violence and harassment is motivated by prohibited factors including, *inter alia*, race. Alternatively, some states have preferred a strategy of sentence enhancement so that the presence of 'racial motivation' leads to a greater criminal penalty. The threat that such measures pose to the constitutional free speech rights of the accused has led to the Supreme Court declaring some hate crimes statutes to be in breach of the First Amendment,[9] leading A. W. Bradley to observe that 'in this area, at least, the United Kingdom and many other countries are fortunate to be without the full rigours of the First Amendment jurisprudence'.[10]

The tension between the State's aims of protecting both minorities and the right to freedom of expression is a long standing issue which will be re-kindled by the new legislation,[11] especially in the light of the new Human Rights Act 1998 which gives effect to the free speech provision in Article 10 of the European Convention on Human Rights.[12] In this note I examine a different challenge for those who seek to regulate racially motivated crime: what does it mean to commit a criminal offence with motivation which demonstrates 'racial hostility'? Issues of motivation and racial hostility are the key to establishing when there is 'racial aggravation' for the purposes of the new legislation. Delineating the meaning of these terms is essential for understanding the role and limits of criminal liability for racist conduct. There are also pragmatic and institutional reasons for prioritising this task. Law enforcement agencies have stressed that their ability to deal with racist crime depends in large part on ensuring that they are presented with precise and

6 The Verbrechensbekampfungsgestz of 28 October 1994 which came into force on 1 December 1994 (BGB1.I 1994, 3186). See Human Rights Watch/Helsinki, *Germany for Germans· Xenophobia and Racist Violence in Germany* (New York: Human Rights Watch, 1995).

7 The Parliament has issued several resolutions and reports on this issue. Early investigation was by special committees. See especially the Report on the Findings of the Committee of Inquiry into the Rise of Fascism and Racism in Europe – The Evrigenis Report, (Luxembourg: December 1985) and the Report of the Findings of the Committee of Inquiry on Racism and Xenophobia – The Ford Report, (Luxembourg, 1991). More recently, the Parliament has passed a number of resolutions on the issue including most recently Resolution B4 – 0045/97 of 30 January 1997 on racism, xenophobia and anti-Semitism and the European Year Against Racism, Official Journal C55, February 1997. The Council has also adopted initiatives including the establishment of a European Monitoring Centre for Racism and Xenophobia, see Council Regulation (EC) No 1035/97, Official Journal L151, 10 June 1997

8 Detailed examples of the US legislation can be found in Lawrence, 'The Punishment of Hate: Toward a Normative Theory of Bias Motivated Crimes' (1994) *Michigan Law Review* 320.

9 See *RAV* v *St Paul* 120 L Ed 2d 305 and *Wisconsin* v *Mitchell* 124 L Ed 2d 436 (1993).

10 'Free Expression and Acts of Racial Hatred' [1992] PL 357, 359 commenting on the decision in *RAV* v *St Paul* above n 9, where the Supreme Court held that Minnesota's hate crimes statutes was in breach of the First Amendment free speech guarantee. For an analysis of the impact of incitement to racial hatred provisions in Britain on the right to free speech see Wolffe, 'Values in Conflict: Incitement to Racial Hatred and the Public Order Act 1986' [1987] PL 85.

11 See Wolffe n 10 above.

12 Human Rights Act 1998 s 1. Strong state controls on racist speech are unlikely to be found to be in breach of Article 10. The decision in *Jersild* v *Denmark* (September 23, 1994) Series A, No. 298) can be explained as an instance where the Court felt that restrictions on the right of a journalist to repeat racist speech fell outside what is justified state action to proscribe racist speech and was therefore in breach of the right to free speech under Article 10. For commentary on the *Jersild* case see R. Wintemute, 'Blasphemy and Incitement to Racial Hatred under the European Convention' (1995–96) 6 KCLJ 143.

© The Modern Law Review Limited 1999

clear definitions of what constitutes 'racially motivated conduct'.[13] Finally, definitions need to draw upon and reflect not only those ideas which find support amongst the victim's (minority) and wider community, but they must also remain sensitive to the beliefs and perceptions which prevail amongst perpetrators and their communities.

The new racially aggravated offences represent an interesting development in the use of the criminal law to protect minorities in the UK. Since the enactment of the Race Relations Act 1976 which treats discriminatory conduct in specified contexts as a statutory tort, the civil law has been adopted as the appropriate paradigm for addressing minority rights issues. The dominance of this model makes it inevitable that the debate on racist crime will be conducted in the shadow of anti-discrimination law. However, any easy assimilation of anti-discrimination law concepts into the law on racially motivated crime should be resisted. Although anti-discrimination law can provide some useful resources for analysis, there are good reasons to keep the issue of racially motivated crime distinct from the more general problems of 'racism' which the Race Relations Act 1976 and other related legislation seeks to address.[14] More specifically, it is argued that the term 'racial aggravation' in the Act should be developed as an autonomous concept by relating it to the underlying problems of racial violence and harassment in the community, rather than through the automatic application of existing anti-discrimination law concepts. Such a strategy has the benefit of ensuring greater clarity and certainty, which is especially important in the context of racist crime. In addition, it is more likely to find support in the wider community, and especially the perpetrator's community, whose co-operation is necessary for ensuring successful enforcement. Defining racial hostility in a way which attends to the causes and contexts within which racist crimes occur in the community is more likely to ensure that the most culpable forms of racist conduct are regulated by the criminal law, whilst at the same time ensuring that the Government's aims of promoting a 'multiracial and multicultural Britain' are realised by the legislation.

Violence and harassment against ethnic minorities

Public concern with racial violence and harassment is not new. Throughout the 1980s policy makers became increasingly aware of the importance of this issue. The action of the Home Office in 1981 which led to the commissioning of the first official study of racial attacks and harassment[15] was prompted by the lobbying of groups such as Joint Committee against Racialism (JCAR) whose influential report was presented to the Home Secretary in February 1981.[16] Since this date government agencies have kept a record of these types of incidents and sought to develop appropriate policy responses. The reason for the renewed interest in this social problem is partly the increasing recognition of the gravity of the harm

13 Maynard and Read: '*Policing Racially Motivated Incidents*', Crime Detection and Prevention Series (Paper 84), (London: Home Office, 1997) 26–27

14 See for example the offence of incitement to racial hatred in section 18 of the Public Order Act 1986 For an overview and discussion of the legislation see Bailey, Harris and Jones, *Civil Liberties: Cases and Materials*, (London: Butterworths, 4th ed, 1995)

15 HC Deb vol 998 col 393 5 February 1981.

16 For a discussion of the historical background see Benjamin Bowling, 'Racial Harassment and the Process of Victimisation' (1993) *British Journal of Criminology* 33. For a discussion of this issue and hate crime in the UK more generally see Hare, 'Legislating Against Hate – The Legal Response to Bias Crimes' (1997) 17 OJLS 415.

© The Modern Law Review Limited 1999

caused by racially aggravated crime to individuals, their communities and to the wider public interest. In addition, the view that racially motivated violence and harassment is a peripheral rather than substantial problem can no longer be sustained in the light of increasingly sophisticated and widely available quantitative data which confirms the exact opposite. Moreover, the assumption that such conduct is ad hoc is being increasingly challenged by evidence which suggests that racist crime in the UK is often the end-product of organised and planned 'political' activity.[17]

The most recent analysis of the 1996 British Crime Survey (BCS) confirms the quantity of incidents during 1995. Its findings distinguish between racially motivated crime and harassment. First in relation to racially motivated crime, it concludes that 143,000 incidents which counted as crimes in the BCS (including threats) committed against black people or Asians were judged – *by them* – to be racially motivated. This represents 15 per cent of all crimes committed against these groups. There was no conclusive evidence that this represented any significant change as compared with the official figures for the 1991 BCS.[18] There was a disparity (partially explained by differences in the definition of what constitutes a racial incident) between the figures in the 1996 BCS and the number of racial incidents recorded by the police. Only 45 per cent of all racially motivated crimes recorded by the BCS were reported to the police. Reporting to the police was less common among ethnic minorities: only 29 per cent of racist crimes where the victim was from an ethnic minority were reported to the police, as compared with 55 per cent where the victim was white. This confirmed the results of the 1988 and 1992 BCS which had concluded that there was under-reporting of racial incidents, and that many incidents with a racial element did not come to police attention.[19]

Under-reporting of incidents and problems with co-operation between victims of racist crime and law enforcement agencies will be a substantial barrier to the success of the new offences,[20] and are bound to remain among the most controversial areas in the debate on regulating racist crime. The categorisation of a crime as 'racial' and prosecution for the new offences will depend on co-operation between the victims of the crime and the police.[21] The pressing need to develop an appropriate institutional framework for dealing with racially aggravated offences was noted by Baroness Amos, who argued that:

> the success of this legislation will rest on the consistency and priority afforded it by key agencies – particularly the police, the Crown Prosecution Service and local authorities. We shall need to ensure that mechanisms are put in place to enable us to judge how well those agencies are doing. This is particularly the case in the light of the findings of the Police Complaints Authority investigation into the death of Stephen Lawrence ... The results of the investigation show how under-performance and a low priority can lead to the mishandling of

17 See Human Rights Watch, *Racist Violence in the United Kingdom*, (Human Rights Watch, USA, 1997) 13–20. See also Satnam Virdee, *Racial Violence and Harassment*, (London: Policy Studies Institute, 1995) Introduction, which discusses the relationship between racist crime and organised political activity. More recently it was disclosed that membership of the far-right group Combat 18 included some soldiers from Britain's elite regiments: *The Independent* 8 March 1999

18 n 1 above.

19 Aye Maung and Miriees Black, *Racially Motivated Crime. a British Survey Analysis*, Research and Planning Unit Paper 82 (London: Home Office, 1994) 20–21.

20 See Maynard and Read n 13 above

21 See David J. Smith, 'The Origins of Black Hostility to the Police' (1991) 2 *Policing and Society* 1 which discusses the reasons for the higher levels of hostility to the police among Afro-Caribbean as compared with white or Asian people. See also Benjamin Bowling, *Violent Racism: Victimisation, Policing and Social Context* (Oxford: Clarendon, 1998).

© The Modern Law Review Limited 1999

cases, and it shows how such failures undermine public confidence – particularly the confidence of ethnic minority communities in the justice system.[22]

Since then continued interest in the relationship between the police and racist crime, sustained by the Stephen Lawrence case, has led to the Home Secretary (Jack Straw) setting out a number of new initiatives aimed at improving the relationship between the police and ethnic minorities.[23]

In relation to harassment, over one in ten people were the victims of harassment (insulted or pestered) in 1995, and for ethnic minorities such harassment was more likely to have a racist element. In total, the 1996 BCS estimated that 200,000 members of ethnic minorities experienced some form of racially motivated victimisation and harassment in 1995.[24] Finally, ethnic minorities scored higher on all the BCS measures of fear of crime: they worry more about falling victim to crime and felt less safe on the streets and in their homes which the 1996 BCS survey concluded was 'to a large extent ... a reflection of their higher risks of victimisation and harassment'.[25]

Quantitative data, although useful for understanding the extent of racist crime, is not a good guide to analysing the causes of racial violence and harassment. The lack of detailed qualitative research in this area has been noted by a number of commentators, who point out that the focus on the incidence of racial violence and harassment fails to take into account the way in which a more dynamic analysis is appropriate.[26] Identifying the causes of this type of harm requires an understanding of the social relationships between all the actors; the continuity across threats, intimidation and violence; repeated victimisation; and the way in which this behaviour is underpinned by intricate inter-relationships between different communities. Although there is clearly a complex set of motivating factors which underpin this type of conduct there is one set of relationships which is especially important to consider if criminalising racist conduct is adopted as part of a general

22 HL Deb vol 584 col 585 16 December 1997. For an anecdotal example of 'underreporting' of racially motivated incidents to the police see the comments of Kirit Gordhandas, a police officer with over ten years experience. He comments on how incidents of racially motivated violence were kept down in one police station in West Yorkshire: 'I was sent to a racially motivated incident and after taking details I returned to the station to complete the racially motivated crime incident book. I was told the book was kept under lock and key so that forms could not be submitted. It was explained to me that this was a way of ensuring that no one submitted such an incident form – thus there were no racially motivated incidents within the division'. See '*When Colour is a Crime*', *The Independent on Sunday*, 21 February 1999, which also discusses the issue of 'institutional racism' in the police force. There have also been a number of comments from senior ranking members of the police force which support the view that racism within the police force is a wide ranging and serious problem. See for example the comments of Dr Robin Oakley, an independent consultant to the police on race, who stated that the problem is not confined to a 'few rotten apples' *The Independent*, 3 August 1998. See also the comments of the Commissioner of the Metropolitan Police, Sir Paul Condon, who stated: 'Is there racism in the police force? Yes, there is. Is it more than a few bad apples? Yes, it is'. *The Times*, 2 February 1999.

23 See his speech to the annual conference of the Black Police Association on 19 October 1998 where he stated that it was not enough to pay lip service to the ethos of equal opportunities and stated that police forces were to be set targets for the recruitment of black and Asian officers: *The Independent*, 29 October 1998.

24 For examples of low level harassment see *Learning in Terror: A survey of racial harassment in schools and colleges in England, Scotland and Wales, 1985–1987* (Commission for Racial Equality: London). Where there is more than one incident of harassment, such low level harassment can now be dealt with under the Protection from Harassment Act 1997 which provides for crimmal sanctions, as well as a remedy in tort for damages or an injunction.

25 n 1 above 1

26 This is a point noted by Bowling who observes: 'Racial victimisation is, like other social processes, dynamic and in a state of constant movement and change, rather than static and fixed. While individual events can be abstracted from this process, fixed in time and place and recorded by individuals and institutions, the process itself is continuous' n 17 above, 238.

© The Modern Law Review Limited 1999

strategy to promote multiculturalism. Sibbit's research on the perpetrators of racial violence and harassment analyses the interplay between perpetrators and the wider community. In his case studies of the perpetrators of racial violence he concludes:

> Based on existing evidence and evidence from two case study areas, it paints a picture which suggests that the perpetrators of the most violent racist assaults do not act in a social vacuum. They carry out their assaults in areas in which all age groups across the local community share common attitudes to ethnic minorities, where people regularly express these views to each other ... regularly engage in the verbal abuse and intimidation of ethnic minorities. ... There therefore appears to be a critical, mutually supportive relationship between the individual perpetrator and the wider community.[27]

Sibbit identifies three ways in which the individual perpetrator is linked to the community: the attitudes of the wider community help to shape the views of individual perpetrators; the actions of extremely violent perpetrators serve as markers by which other perpetrators are able to judge that their own abusive and intimidatory actions are relatively harmless; and the perpetrator serves the wider community in a vicarious manner by taking their collective views and acting them out. The research makes clear that any strategy for dealing with racial violence and harassment will need to focus on not only the individual conduct of the perpetrator, but also the relationship between this conduct and the wider community which may support and articulate the beliefs which underpin racist crime.

A number of other conclusions are noteworthy. Sibbit suggests that any attempt to reduce or prevent racial violence and harassment needs to distinguish between three sets of strategies: first, identification and effective action against perpetrators; second, identifying potential perpetrators and the development of effective strategies to prevent them becoming actual perpetrators; and finally, the development of a range of strategies for consistently addressing the perpetrator community's general attitudes towards ethnic minorities.[28] Sibbit notes that there is a significant role for non-legal agencies with responsibility for tackling racism within perpetrator communities. He concludes:

> It is in this area of work that there may be some cause for optimism. A number of examples discovered in the case studies – and elsewhere – have demonstrated that individuals within perpetrator communities and perpetrators themselves have, on occasions, transcended (or side-stepped) the local tendency to see others in purely racial terms. It is where there has been space for individual relationships to develop that – even within a seemingly rigid perpetrator community – relations have been able to cross apparent boundaries of ethnicity.[29]

Legal regulation of racial hostility

Traditionally, offences involving racial hatred have been characterised as 'public order' problems in the United Kingdom.[30] An offence of 'incitement to commit racial hatred' was introduced by the Race Relations Act 1965 by means of an amendment to the Public Order Act 1936. Its goals have always uncomfortably straddled the objectives of public order on the one hand and equality as reflected in

27 Rae Sibbit: *The Perpetrators of Racial Harassment and Racial Violence*, Home Office Research Study 176 (London: Home Office, 1997), 11–18 See too chapters 4 and 5.

28 *ibid* 106.

29 *ibid* 109

30 See Lacey, Wells and Meure, *Reconstructing Criminal Law* (London: Weidenfield & Nicolson, 1990) 118–120.

© The Modern Law Review Limited 1999

the Race Relations Act 1976 on the other. The Public Order Act 1986 reformed the offence of incitement to racial hatred,[31] as well as introducing a new offence of possession of racially inflammatory material and causing harassment, harm and distress.[32] The Government has built on 'sentence enhancement' for racially aggravated crimes as its preferred strategy for dealing with racial violence and harassment, thereby following the practice preferred in a number of American state legislatures.[33] Under sections 29–32 of the Crime and Disorder Act 1998, where the prosecution can prove that the relevant 'basic offence' was 'racially aggravated' there is a mandatory requirement to add to the existing range of the sentence through additional fines and periods of imprisonment which are specified in the Act. This sentence enhancement means that the basic offence is transformed into a 'racially aggravated offence'. The 'basic offences' include: common assault; assault occasioning actual body harm and malicious wounding under the Offences against the Person Act 1861; criminal damage; offences under the Public Order Act 1986 and offences under the Protection from Harassment Act 1997.[34]

Judicial sentencing discretion, which allows a judge to take into account 'racial motivation' as a factor in determining the final sentence, has been an effective weapon in dealing with racially motivated crime. In this way, even before the introduction of the racially aggravated offences under the Act there was some potential for dealing with racially motivated crime by imposing a more severe sentence. The case of *R* v *Ribbans, R* v *Duggan, R* v *Ridley*[35] affirmed this principle in the strongest terms. Following that case where racial motivation was proved the Court was required to take this factor into account at the sentencing stage as an aggravating feature. Lord Chief Justice Taylor stated: 'We take the view that it is perfectly possible for the court to deal with any offence of violence which has a proven racial element in it, in a way which makes clear that that aspect invests the offence with added gravity and therefore must be regarded as an aggravating feature'.[36] However, this does not end the argument in favour of specific legislation to deal with racial violence and harassment. In particular, it has been persuasively argued that 'the most serious objection to treating racist motivation as just one more factor to be weighed in the aggravation/mitigation calculus is that to do so fails to offer a sufficient denunciation of the conduct in question'.[37] In *R* v *Craney and*

31 s 18. See too the Football (Offences) Act 1991 s 3 which makes it an offence to take part in racist chanting at a designated football match

32 s 5. For a discussion of these provisions see Feldman: *Civil Liberties and Human Rights in England and Wales* (Oxford. Clarendon Press, 1993) 812–814.

33 See Lawrence n 8 above for detailed examples of sentence enhancement statutes. The provision of the Wisconsin criminal code which passed the scrutiny of the Supreme Court in *Wisconsin* v *Mitchell* (n 9 above) as compatible with the accused's First Amendment free speech rights is a good example. The relevant provision enhances the maximum penalty for an offence whenever the defendant 'intentionally selects the person against whom the crime . is committed ... because of the race, religion, colour, disability, sexual orientation, national origin or ancestry of that person'.

34 The full details are as follows. Racially aggravated assaults (Crime and Disorder Act 1998 s 29): basic offence is an offence under the Offences against the Persons Act 1861 s 20 (malicious wounding or grievous bodily harm); or s 47 (actual bodily harm), or common assault; racially aggravated criminal damage (Crime and Disorder Act 1998 s 30): basic offence is an offence under the Criminal Damage Act 1971 s 1 (1) (destroying or damaging property belonging to another); racially aggravated public order (Crime and Disorder Act 1998 s 31): basic offence is an offence under section 4 (fear or provocation of violence), 4A (intentional harassment, alarm or distress) or 5 (harassment, alarm or distress) of the Public Order Act 1986. Racially aggravated harassment (Crime and Disorder Act s 32)· basic offence is an offence under the Protection from Harassment Act 1997 s 2 (offence of harassment) or s 4 (putting people in fear of violence).

35 (1995) Cr App R (S) 698.

36 *ibid* 702.

37 Hare n 16 above 423.

© The Modern Law Review Limited 1999

Corbett[38] for example, the Court refused to give detailed guidance on the extent to which racial motivation aggravated the offence, suggesting that it should be treated as one of a number of different factors taken into the balance rather than a factor which has special weight during the sentencing process. Section 82 of the Act gives statutory effect to the decision in *R* v *Ribbans* and applies to all offences other than those deemed to be basic offences in Part II of the Act. Where there is evidence of racial aggravation as part of the offence, the judge is required to consider this and impose a higher sentence within the maximum available. In addition, section 82(2) makes it clear that the court will state in open court that the offence is racially motivated, thereby ensuring that the rhetorical power of sentence enhancement is conveyed to the victims, perpetrators and the community.

Section 28 of the Act sets out the test which is to be applied in determining whether an offence listed under sections 29 to 32 (the basic offences) is racially aggravated. It is also the test which the court applies in considering an increase in sentence for offences other than those under sections 29–32 (see section 82). Section 28 introduces a test for 'racially aggravated' which uses two sets of definitions. First, section 28(1)(a) states that there is racial aggravation where 'at the time of committing the offence, or immediately before or after doing so, the offender demonstrates towards the victim of the offence hostility based on the victim's membership (or presumed membership) of a racial group'. Second, section 28(1)(b) adds to this meaning by confirming that racial aggravation is also established where 'the offence is motivated (wholly or partly) by hostility towards members of a racial group based on their membership of that group'. In addition, section 28(3)(a) states that it is immaterial that the offender's hostility is also based on either 'the fact or presumption that any person or group of persons belongs to any religious group' or any factor not mentioned in that paragraph. This last provision confirms that 'racial hostility' need not be the sole motivation for commission of the basic offence.

The coverage of racially aggravated offences

Before moving on to consider the definition of racial aggravation in section 28 of the Act, it is worth noting the potential problems with the coverage of the racially aggravated offences. The explanatory provision in section 28(2) confirms that 'membership' in relation to a racial group, includes association with members of that group and 'presumed' means presumed by the offender. Section 28(4) sets out the definition of the term 'racial group'. It means 'a groups of persons defined by reference to race, colour, nationality (including citizenship) or ethnic or national origins'. This is the same as the definition of 'racial group' in section 3(1) of the Race Relations Act 1976, in relation to which the term 'racial' was given an extended meaning by the House of Lords in *Mandla* v *Dowell Lee*.[39] In this way discrimination law has provided concepts which have been explicitly 'borrowed' for the purposes of the racially aggravated offences. This was also the preferred strategy for defining 'racial group' for the purpose of the criminal offence of incitement to racial hatred in Part III of the Public Order Act 1986.

However, the strategy of using the Race Relations Act 1976 as the alpha and omega for the legal regulation of violence and harassment against vulnerable

38 (1996) 2 Cr App R(S) 336. For comment see [1996] Crim LR 525
39 [1983] 2 AC 548.

© The Modern Law Review Limited 1999

minority groups is not ideal. It fails to take into account the compound and overlapping nature of the characteristics of these minorities which in turn provoke hostility. In the House of Commons debates on the Bill this definition was the subject of heated debate. The Government was asked why the existing legislation failed to protect two sets of minority groups: non-ethnic religious minorities[40]; and gays and lesbians.[41] The Government resisted pressure to extend protection to groups other than 'ethnic' groups, but in the face of strong questioning on this issue, it agreed to re-consider its position in the light of future developments and research. However, it justified the priority given to race and ethnicity by stating that these are especially important to its goal of multiculturalism. It is this last goal which was cited by the Government as the main justification for prioritising violence and harassment against 'ethnic minorities'. Echoing the Labour Party's manifesto pledge, the Home Office Minister Mike O'Brien outlined the government's thinking by stating: 'The Government are tightly focused on tackling racist crime ... Racial attacks ... tear a hole in the fabric of our multiracial society and the Government want to send a clear, specific and focused message about racial motivation, which is why we included it in the Bill'.[42] In the light of evidence that there is a real risk of violence and harassment against religious groups and gays and lesbians,[43] it will be interesting to see how long the strategy of distinguishing between different types of minority groups will be maintained.[44]

Even on the Government's own terms – of prioritising racist crime because of a commitment to multiculturalism – the present strategy of borrowing concepts from the Race Relations Act 1976 is problematic. Criticism has focused on the way in which the term 'racial group' arbitrarily requires a distinction to be drawn between the various overlapping factors which go to determine the 'ethnicity'[45] of the group. In particular, there are problems in maintaining a clear demarcation between on the one hand, ethnicity, race, colour and culture; and on the other religion, which is specifically excluded from the definition of 'racial group' in the Act. Policing the boundaries between these overlapping characteristics has caused considerable problems in the context of the Race Relations Act 1976 which extends protection to some religious groups because they also qualify as a racial group, but does not extend coverage to all religious groups in a consistent or coherent way.[46] This is an especially important point in regulating racial violence

40 HC Standing Committee B col 312 12 May 1998; Mr Clappison moved an amendment that religion as well as race should be covered by the legislation. See also HC Deb vol 314 23 June col 903 1998 where the Home Secretary Jack Straw stated that 'we of course recognise the case for laws dealing with religious discrimination and hatred, we are giving it active consideration but we need to give such offences the same careful consideration that we have given to these offences.'

41 HC Standing Committee B cols 840–854 11 June 1998.

42 *ibid* col 848

43 Attacks on Jewish cemetries have taken place in Leeds, Manchester and Edmonton (see *Daily Mail* 5 August and 26 August 1990; *The Times* and *Independent* 30 July 1990). Twenty mosques were fire-bombed during the Gulf conflict (see *Guardian* 6 February 1991 and *The Times* 17 January 1991). See also *Labour Research* (October 1991) Vol 80, no 10, 15 which concluded that there had been an increase in the levels of racial violence specifically targeting Muslims during the Gulf War. See also Mason and Palmer, *Queer Bashing – A National Survey of Hate Crimes Against Lesbians and Gay Men* (London: Stonewall, 1993).

44 The alternative is to include a wider and more general list of characteristics which trigger the enhanced penalty provisions. For an American example see the *Wisconsin* v *Mitchell* statute n 9 above, discussed by Lawrence n 8 above.

45 For a discussion of the difficulties raised by defining the term ethnicity see Keller, 'Re-thinking Ethnic and Cultural Rights in Europe' (1998) 18 OJLS 29, 34–43.

46 Sikhs and Jews are covered by the term racial group, see *Mandla* v *Dowell Lee* n 39 above and *Siede* v *Gillette Industries* [1980] IRLR 42. However, there are problems and considerable confusion over the status of other minority groups who can be easily identified because of their religious observance and

© The Modern Law Review Limited 1999

and harassment in the light of evidence which suggests that such conduct is often specifically aimed at ethnic minorities because of characteristics which link them to a particular religious group.[47] It is likely that in most cases (with the exception of non-ethnic religious minorities) which are motivated by hostility based on religion there will be an overlap with the categories of ethnicity and culture and the case will therefore fall within the definition of racial group under the Act. Under the Race Relations Act 1976, the prohibition of indirect discrimination ensures that the application of criteria on the grounds of religion, which have a disproportionate impact on a particular racial group, will fall within the definition of racial discrimination. Where there is an overlap of racial and religious motivation, section 28(3) of the 1998 Act makes clear that the test for racial aggravation set out in section 28(1)(a) and (b) requires racial hostility to be 'wholly or partly' a motivating factor. Thus hostility which is mixed (partly racial and partly based on religion) will be covered by the Act.

However, there are likely to be problems with such a piecemeal and arbitrary approach to an issue as critical as the coverage of the racially aggravated offences. Home Office research supports the view that terms such as Asian are too crude to capture the diversity of factors which will be relevant to understanding the motivation of the offender and the feelings of the victim.[48] The racial offence is seen to be more serious than the basic offence because of the additional harm to the individual and the group. By forcing an attack motivated by one type of individual characteristic (such as religion) into the category of 'ethnicity' or 'culture', there is the danger of trying to fit a square peg into a round hole.[49] The individual victim and the relevant community group will not feel that the harm they have suffered has been recognised by the law. Given that attacks against ethnic minorities are often a prelude to more serious and widespread racial tensions, it is important that the new offences are perceived as an adequate response to grievances, thereby alleviating the fears and resentment of individuals and groups. In this context, failing to ensure that the definition of the harm suffered by victims (individuals and groups) correlates with legal definitions and remedies may have a detrimental impact on good race relations.

A further consequence of this definition of 'racial group', and potential gap in the coverage of the Act, is that organised groups and persistent offenders will modify their behaviour in response to the patchy coverage of the new criminal offence. They may prefer to choose religious institutions as targets, focus on certain individuals who are unlikely to be unprotected by the legislation (eg white, English converts to a religion) and to use language (eg which very specifically focuses on religion rather than race) in order to avoid the risk of being liable for the more serious 'racially aggravated offence'. For example, it may be to the advantage of a perpetrator to desecrate a Sikh temple or a Roman Catholic church frequented by Irish people rather than a church frequented by a mixed congregation or a mosque. It will be a better strategy to attack a Rastafarian with dreadlocks or a

are therefore potentially high risk groups, for example, Rastafarians (see *Dawkins* v *Department of Environment* [1993] IRLR 284) and Muslims (see the cases cited in IDS Brief 538, April 1995). Further, as Muslims did not fall under the category 'racial', the CPS decision not to prosecute a branch organiser of the British National Party using 'Rights for Whites' stickers in relation to the conversion of an old dairy into a mosque under the incitement to racial hatred provisions could not be challenged by the London Borough of Merton (See *The Guardian*, 28 October 1998).

47 See n 43 above.

48 *Ethnic Minorities: victimisation and racial harassment; Findings from the 1988 and 1992 British Crime Surveys* Home Office Research Study Paper 154, Home Office Research Studies 1996, 55.

49 See Modood, *Racial Equality: Colour, Culture and Justice*, The Commission on Social Justice, IPPR, 1994.

© The Modern Law Review Limited 1999

Muslim with a beard rather than a Jewish man wearing a kippah or a Sikh man wearing a turban. Leaving aside the obvious arguments of principle against such arbitrary distinctions, there will be confusion in attempting to monitor the demarcation boundaries between these overlapping characteristics which trigger the hostility of perpetrators against minorities. This is bound to add to the existing difficulties in proving racial hostility and motivation. As noted earlier it is a characteristic of violence against minorities in the UK that it consistently takes the form of organised and orchestrated activity which specifically targets certain vulnerable groups.[50] Therefore, there is a real risk of adaptation to these anomalies in the coverage of the racially aggravated offences which could result in their being circumvented by 'clever' and organised perpetrators. If this happens, the legislation may leave wholly unregulated some of the most extreme forms of violence against minorities.

Racial hostility

Section 28 of the Act makes it clear that it is the presence of 'racial hostility' that is the key triggering device which transforms the basic offences into the racially aggravated offences: assault (section 29); criminal damage (section 30); public order offences (section 31) and harassment (section 31). How do the racially aggravated offences relate to the underlying basic offences? The new offences shift issues of the motivation of the perpetrator which are characteristically reserved for sentencing back into the definition of the racially aggravated offence. The main focus for any justification for these offences will be on the greater culpability of the perpetrator who acts out of 'racial hostility'. However, this emphasis on fault should not obscure the fact that the racially aggravated offences are not merely an instance of the same harm (of the basic offences) being carried out in a certain serious way (with racial aggravation). The new racially aggravated offences are aimed at conduct which causes a harm of a *qualitatively* different type to that caused by the basic offences. Delineating with greater precision the nature of this harm should be a high priority for those analysing racially aggravated offences, thereby enabling a more principled development of the new offences. Here, it is worth noting that this project will have to focus on the harm of racially aggravated offences at a number of different levels. At an individual level, where a victim is targeted because of a characteristic such as race, this is an additional harm of a greater and of a more serious order than a 'normal' incident of violence and harassment. In addition, the harm to the victim's community which is said to experience the harm of 'racial hostility' in a vicarious way is also relevant.[51] Finally, racist crime has the potential to undermine the State interest in good race relations, and it is this last goal which is cited by the Government as its main justification for prioritising violence and harassment against 'ethnic minorities' as part of its aim to promote multiculturalism.

There is a clear and obvious relationship between the new racially aggravated offences and the justification for anti-discrimination law generally, and the Race

50 see n 17 above.

51 See S. Virdee, *Racial Violence and Harassment* (London: Policy Studies Institute 1995) ch 4 which discusses the effects and constraints on the lifestyles of ethnic minorities caused by the fear of racial violence and harassment. See also the discussion of the psychological effects of this type of conduct in Joan C Weiss 'Ethnoviolence: Impact Upon and Response of Victims and the Community', in Robert J. Kelley (ed), *Bias Crime: American Law Enforcement and Legal Responses* (Chicago: University of Illinois, 1993).

© The Modern Law Review Limited 1999

Relations Act 1976 in particular. Both are based on the same idea: that acts and conduct which express contempt for members of a minority on the basis of their group characteristic are a serious harm. Those seeking to understand the new racially aggravated offences can therefore usefully draw on a substantial body of existing analysis of the anti-discrimination principle generally, and race discrimination in particular.[52] However, the continuity between the two areas of law also poses dangers. As stated, not only the additional harm but also the greater fault of the perpetrator provides the justification for differentiating between the 'ordinary' basic offences and the more heinous racially aggravated offences. There are a number of stages which need to be kept separate in analysing the state of mind of a perpetrator of the new racially aggravated offences. It needs to be established that he had the *mens rea* which is required to complete the commission of the basic offence, for example intention or recklessness in the case of racially aggravated assaults under section 29 of the Act. In addition, the prosecution needs to prove beyond reasonable doubt that (a) 'at the time of committing the offence, or immediately before or after doing so, the offender demonstrates towards the victim of the offence hostility based on the victim's membership (or presumed membership) of a racial group' or (b) the offence is 'motivated (wholly or partly) by hostility towards members of a racial group based on their membership of that group'. How should the requirement for 'racial hostility' be construed in this context?

It is at this stage that it is tempting to borrow from the developed concepts of direct (and to a lesser degree indirect) discrimination. Such an enquiry encourages analysing racist conduct by examining its impact on the victim and the community. The focus is on whether from the perspective of the victim or the wider community there is evidence which demonstrates the racial hostility or motivation of the perpetrator. If 'racial hostility' is being understood from the perspective of the victim or the community, then any further enquiry about the state of mind of the perpetrator will be irrelevant. Such an approach would be consistent with developments in anti-discrimination law. The interpretation of direct discrimination which is relevant for both the Sex Discrimination Act 1975 and the Race Relations Act 1976 has shifted from a test which focuses on the subjective intention of the discriminator[53] to one which is a more objective test for liability. In *R* v *Birmingham City Council, ex parte Equal Opportunities Commission*[54] and *James* v *Eastleigh BC*,[55] Lord Goff considered the construction of the phrase 'on the ground of her sex' in section 1 (1)(a) of the Sex Discrimination Act 1975. In *Birmingham* Lord Goff stated: 'The intention or motive of the defendant to discriminate ... is not a necessary condition of liability; it is perfectly possible to imagine cases where the defendant had no such motive, and yet did in fact discriminate on the ground of sex'.[56] This move towards a more objective test was confirmed in *James* with the result that issues of liability for the purposes of discrimination focus more closely on results than issues of intention. Lord Bridge affirmed this in his judgement in *James* where he stated: 'the purity of the discriminator's subjective motive, intention or reason for discriminating cannot save the criterion applied from the objective taint of discrimination'.[57]

52 For an overview of the literature see McCrudden (ed), *Anti-Discrimination Law* (Aldershot: Dartmouth, 1991) especially the Introduction. See also Gardner, 'Liberals and Unlawful Discrimination' (1989) 9 OJLS 1.

53 See for example *Ministry of Defence* v *Jeremiah* [1980] 1 QB 87; *Gill* v *El Vino Co* [1983] QB 425.

54 [1989] 1 AC 1155.

55 [1990] 2 AC 751

56 n 54 above 1194.

57 n 55 above 765–766.

© The Modern Law Review Limited 1999

Although this move towards a more objective, results-focused approach to the concept of discrimination has not been unproblematic in anti-discrimination law,[58] it does demonstrate an increasing tendency on the part of the courts to adopt a strong response to the problems of discrimination in the civil law. Is a similarly trenchant and results focused model appropriate in the context of the criminalising racist conduct? It is suggested that such a model would be inappropriate, and that the objective proof of discrimination along these lines is an insufficient guide to the presence of racial hostility and motivation as required by section 28 of the Act. For example, where a perpetrator of one of the basic offences has selected the victim 'because of her race', this can be cited as evidence of racial hostility. Rather than being an anonymous act of random selection of the victim, the fact that the victim has been chosen because of membership of a racial group will reveal hatred. However, discriminatory selection should not be taken as conclusive proof of racial motivation in this context. An offender may choose to attack a member of a racial group – such as an Asian woman – because (rather like the choice of a victim because of his or her age or disability) she is less likely to respond and resist. In this example of discriminatory selection on the grounds of race, the motivation may itself be based on prejudices that Asian women are more passive and less likely to resort to anger or retaliation. However, this type of motivation should not of itself and automatically be sufficient to translate the basic offence into the racially aggravated offence.

Stereotyping in this way, is not synonymous with the racial hostility which is required by section 28 of the Act. Civil anti-discrimination law may legitimately be concerned with the regulation of the manifestations of unconscious stereotypes, but for the purposes of the criminal law the perpetrator must have been subjectively aware that his conduct was the manifestation of racial hatred.[59] An act based on stereotyping does not express or convey the idea that the victim is held in contempt because of membership of a particular racial group. In discrimination law, the reasons for the discriminator selecting the victim 'because of' membership of a racial group may be irrelevant; even if benign they will be caught by the definition of discrimination under the Race Relations Act 1976. However, in the context of racist crime, there needs to be the attribution of moral culpability to the perpetrator to justify the greater seriousness and higher penalties of the racially aggravated offences. Racial hostility is required, and makes these crimes more heinous, not because it indicates the type of faulty or inflexible generalisation towards an individual which characterises racial stereotyping, but rather because it is evidence of a conscious animus which is deliberately directed towards an individual because of his or her membership of a protected group. Discriminatory selection of a victim should not therefore be a sufficient alternative for the proof of such animus as the essential element which transforms the basic offence into the racially aggravated offence. Therefore, although the Race Relations Act 1976 and the 1998 Act will often overlap, discriminatory selection of a victim may often fall outside the definition of racial hostility under section 28 of the 1998 Act. In such circumstances, it may be more appropriate to use strategies other than the criminal law to challenge attitudes and stereotypes within the perpetrator communities and in the media and wider public culture.

58 See the discussion in Bamforth: 'The Changing Concept of Sex Discrimination' (1993) 56 MLR 872. For a discussion of the 'but for' test and *James* see Gardner, 'On the Grounds of Her Sex(uality)?' (1998) 18 OJLS 167.

59 See Ashworth: *Principles of Criminal Law*, (Oxford: Clarendon, 1995) 83–85. See too Sibbit, n 27 above, who suggests that it is possible to work through and re-educate in relation to these stereotypes.

© The Modern Law Review Limited 1999

Subjectivism is especially important in these contexts in the light of evidence which suggests that racist attitudes and conduct are prevalent within the perpetrator community and are often automatically assimilated and articulated. Sibbit has noted that racial harassment and violence occurs in communities where 'people of all ages, including very young children and older adults, regularly engage in verbal abuse and intimidation of ethnic minorities'.[60] In this context it is not difficult to envisage a scenario of negligent racist conduct which should not lead to criminal liability for the racially aggravated offences. A perpetrator may merely 'parrot' and repeat a racist term in the course of carrying out an attack, or use an offensive sign or means of carrying out an attack without being aware of the full implications of the use of such symbols. For example, criminal damage may be caused by a perpetrator in a whole range of ways. What if the perpetrator uses language ('Wogs Out') or a symbol (a swastika) which objectively and according to its interpretation by the victim and the community is evidence of racial hostility? In these circumstances, if this conduct is being judged according to its impact on the community or by an objective test then this is a clear case of racial hostility. If a subjective standard is used the prosecution needs to show that the perpetrator intended or was knowingly reckless as to the use of this language or these symbols. As a practical matter, the use of such obviously racist language and symbols will be the strongest evidence that there is such subjective racial hostility, but there may be instances where the use of racist language may be coincidental or serve another function. Hare cites examples of instances where the use of racist language may be coincidental or serve the function of identifying the victim rather than a conscious insult which demonstrates racial hostility and deliberately singles out the victim for contempt.[61] This point becomes clearer by focusing on the use of more subtle means which could be construed as a demonstration of racial hostility, for example, if a perpetrator accidentally used pig meat to attack a Jewish person, or alcohol to attack a Bangladeshi person, or cow's leather to attack someone who is Indian. In these cases, from an objective reasonableness test, or using the 'impact on the victim or community' criteria, this will undoubtedly be an instance where there is a demonstration of racial hostility. However, this is exactly the type of case where the application of subjective criteria will provide the perpetrator with an opportunity to argue that his use of this mode for the attack (although it may be negligent conduct) was not a demonstration of his or her racial hostility.

In this context it could be argued that there is no necessity for such 'correspondence' between the perpetrator's state of mind and the harm; that there is no need for the intention or recklessness to be related to the particular harm of racial hostility.[62] The fact that the perpetrator has carried out the basic offence, and has done so in a way which could be taken as evidence of racial hostility, means that he or she exposes himself or herself to the risk that he or she will be liable for the more serious racially aggravated offence. Such a response is especially tempting in the context of racially aggravated crime where proving racial hostility is notoriously difficult, and where these difficulties have led some commentators in

60 n 27 above 100 and ch 4 'Constructing a profile of perpetrators' which confirms that many of those who articulate racist views are young and with low educational qualifications. Recent Home Office Studies confirm that the perpetrators of low level racial harassment are often found to be children between four and ten. The Home Office is looking into the use of the Child Safety Order and Parenting Order to prevent acts of harassment by children: *The Independent on Sunday* 21 June 1998.

61 n 33 above 439, where he providers a practical example of the distinction between the use of racist language and racist motivation in relation to *R* v *Cole (and others)* [1993] Crim LR 300.

62 See Ashworth n 59 above 85.

© The Modern Law Review Limited 1999

the USA to argue that less stringent standards of proof are necessary if the criminal law is to be successful in addressing the problems of racist crime.[63]

Although this may be justified in the context of other criminal law provisions, there are good arguments to avoid such an approach to the new racially aggravated offences. Any policy of criminalising racist conduct has to perform a delicate balancing exercise. On the one hand, it has to express the idea that acts which express contempt for individuals on the grounds of their membership of an ethnic group are a serious harm which justifies the strongest action. However, this aim has to take into account the perceptions of communities which are most likely to be affected by this legislation. Although always an important constraint, it is particularly important that the new racially aggravated offences should reflect attitudes and meanings which are likely to find support in the wider community, and especially the communities where racist crime occurs.

Research suggests that within these communities there exists a continuum between 'acceptable' racist views and stereotypes and the violent behaviour of the perpetrators of serious racist crime.[64] In relation to the most serious cases, the criminal law serves an important specific function which goes beyond its impact on the individual perpetrator, the victim and community interests. Criminalising the most serious forms of racist conduct sends a clear message to potential perpetrators and supporters within the perpetrator community, many of whom set their own standards of acceptable 'racist conduct' by using more serious attacks as a benchmark, that these types of acts are heinous and will lead to additional criminal sanctions.

Yet paradoxically, prevailing racist attitudes also set important constraints on the use of the criminal law in these contexts. The success of any policy against racial violence and harassment needs to stay attentive to the attitudes and beliefs of perpetrator communities where racist attitudes prevail. Legal definitions and strategies must draw on and reflect distinctions which make sense and have meaning for the social actors. In this context, the insistence on subjectivism ensures that only the most serious forms of racist conduct, which are without doubt an expression of conscious racial hatred and not just prejudice or discriminatory attitudes, are regulated by the criminal law. This is more likely to reflect distinctions which make sense to the relevant parties within the perpetrator communities: that those who commit acts of violence and harassment of minorities out of a deliberate and conscious hatred based on their membership of a racial group deserve to be penalised by the criminal law. Of course, this does not mean that other forms of racist conduct are acceptable or that legal and policy responses other than the use of the criminal law should not be actively adopted.

Ensuring a high degree of willing compliance and co-operation with the policy underlying the new racially aggravated offences is also important from the perspective of victims and their communities. Where there is a wide divergence between concepts and definitions which are developed in the law and those which are used in communities, it is likely that there will be a build up of resentment and a 'backlash' against minorities who may be viewed as enjoying an 'unfair advantage' under the criminal law. The fact that ethnic minorities are concentrated in the lowest income groups in the economy means that they will have to share and

63 Gardner, 'Rationality and the Rule of Law in Offences Against the Person' (1994) CLJ 502. For a discussion of the American debate on shifting the burden of proof in hate crimes cases see J. Morsch, 'The Problem of Motive in Hate Crimes: The Argument Against Presumptions of Racial Motivation' (1991) 82 J Crim L and Criminology 659.

64 Sibbit n 27 above ch 3

© The Modern Law Review Limited 1999

compete for scarce public goods (local housing, schools and other public facilities) with potential perpetrators on a daily basis.[65] These individuals have to continue to function and co-exist in communities where racist attitudes prevail and where the perpetrators of racist crime often have wide support. Any policy of singling out certain groups for special protection under the criminal law which fails to attend to the views which prevail in these social contexts is likely to place at risk exactly those individuals and communities which the new racially aggravated offences seek to protect.

Conclusion

The criminal law has a role to play in addressing the most extreme consequences of racism.[66] It is a mistake however to over-estimate this role. Any strategy which is too ambitious and which seeks to be tough on racism by extending the definition of racist crime is likely to undermine the interests of those whom it seeks to protect. The use of the criminal law should be reserved for the most serious and culpable instances of racist conduct. There are other strategies available, both civil law and non-legal measures which require public funding, for those who fall outside this narrow category. Clear and meaningful distinctions need to be drawn and maintained between the behaviour of those who undertake criminal acts motivated by conscious racial hatred on the one hand, and those who are potential perpetrators or passive supporters. The Government's aim of protecting ethnic minorities from intimidation and promoting multiculturalism requires a nuanced and complex strategy. The introduction of racially aggravated offences can make a contribution towards this goal, but it cannot be a substitute for it.

65 See Modood, *Ethnic Minorities in Britain: Diversity and Disadvantage*, (London: Policy Studies Institute, 1997).

66 For the view that criminal statutory regulation does not have a significant role in regulating racist behaviour see F. Brennan, 'Racially Motivated Crime: the Response of the Criminal Justice System' (1999) Crim LR 17.

© The Modern Law Review Limited 1999

CASES

Negligent Misstatement, Threats and the Scope of the *Hedley Byrne* Principle

Richard Mullender*

In recent years, senior members of the judiciary have emphasised the breadth of the principle in *Hedley Byrne & Co Ltd* v *Heller & Partners Ltd* which, of course, makes negligent misstatements actionable.[1] *Hedley Byrne's* breadth was, for example, a prominent theme in Lord Goff's speech in *Henderson* v *Merrett Syndicates Ltd*.[2] His Lordship's words have recently been rehearsed by the Court of Appeal in *Welton and Another* v *North Cornwall District Council*.[3] In this case, a unanimous Court has held that the *Hedley Byrne* principle is broad enough to encompass threats issued by an environmental health officer employed by a local authority.

In one respect, the Court of Appeal's decision appears cautious. The liability rule established in *Welton* is confined to fact-situations of the sort on which the plaintiffs' claim was based. This point notwithstanding, the decision reached by the Court can be regarded as bold. Hitherto, threats have not been identified as falling within the *Hedley Byrne* principle. Both the reasons for regarding the Court's decision as cautious and those for viewing it as bold are examined in this piece. Further, the decision is criticised on three grounds. First, the Court's claim to have specified a liability rule that falls squarely within *Hedley Byrne* is identified as unconvincing. Secondly, it is argued that the Court has responded to *Welton* in a manner that has undermined the integrity of the *Hedley Byrne* principle. Thirdly, an argument is advanced in support of the view that the claim in *Welton* should have been identified as disclosing a novel form of actionable wrong, uncomprehended by existing heads of liability.

The facts and first instance decision

The plaintiffs ran a guest house in Cornwall. Guest houses constitute 'food premises'. Hence, they are subject to a (local authority-administered) statutory scheme of regulation under the Food and Safety Act 1990. This regulatory scheme has the purpose of promoting the public interest in hygienic food preparation. To this end, environmental health officers can, *inter alia*, inspect food premises. In October 1990, one such officer, Mr Brian Evans, called at the plaintiffs' guest

* Law School, University of Newcastle-upon-Tyne.

1 [1964] AC 465.

2 [1995] 2 AC 145, 178–181. (Lord Goff's speech was approved by his four colleagues.) *Hedley Byrne's* breadth has recently been emphasised by the House of Lords in *Williams and Another* v *Natural Life Health Foods Ltd and Another* [1998] 2 All ER 577, 581–582, *per* Lord Steyn. (Lord Steyn's four colleagues approved his speech.)

3 [1997] 1 WLR 570.

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA.

house.[4] He 'conduct[ed] the first ever local authority inspection of the [plaintiffs'] premises' on behalf of his employers, North Cornwall District Council.[5] Mr Evans undertook this inspection in order to determine whether the plaintiffs' premises met the requirements of a recently promulgated statutory instrument.[6]

During his inspection, Mr Evans identified thirteen alterations the plaintiffs would have to make to their premises in order to comply with the law.[7] He described these alterations as 'requirements' and threatened to close down the plaintiffs' business if they were not made.[8] In the following months, Mr Evans made a number of visits to the plaintiffs' premises to inspect and approve the work in progress. The work involved the making of 'major alterations' to the guest house's kitchen.[9] Having made these alterations, the plaintiffs discovered that many of the requirements imposed by Mr Evans fell 'outwith the legislation' under which he purported to act.[10] Hence, they began proceedings in negligence against North Cornwall District Council. They claimed damages for the unnecessary expenditure incurred by them as a result of their acting as specified by Mr Evans.

At trial, the plaintiffs argued that liability should be imposed on the defendant local authority under the *Hedley Byrne* principle. This argument was accepted by Judge Anthony Thompson QC.[11] The judge did not regard the fact that the plaintiffs had been subjected to 'compulsion' by Mr Evans as precluding a claim based on *Hedley Byrne*. He was satisfied that their claim fell within the *Hedley Byrne* principle since four conditions that provide a secure basis upon which to advance a negligent misstatement claim could be satisfied. First, the advice was given for a particular purpose. Secondly, the advice was communicated directly to the plaintiffs. Thirdly, Mr Evans knew that the plaintiffs would rely on his advice. Fourthly, the plaintiffs suffered a detriment as a consequence of relying on the advice given by Evans.[12] Following this decision, the defendants appealed, arguing that no duty of care was owed by Mr Evans.

The decision of the Court of Appeal

Many of the points made by Judge Thompson were rehearsed by the Court of Appeal. The plaintiffs had relied on Mr Evans.[13] Evans knew that they were relying on him.[14] The detriment suffered by the plaintiffs resulted from their taking the action indicated by Evans.[15] Moreover, the three Appeal Court judges rejected the argument that the plaintiffs could not found their claim on *Hedley Byrne* since Mr Evans had issued threats and not simply provided information and/or advice.[16] The

4 The power to make this inspection was conferred on Mr Evans by the Food Act 1984. This Act was superseded (as from 1 January 1991) by the Act of 1990. See *ibid*, 575, *per* Rose LJ.
5 *ibid*, 573, *per* Rose LJ.
6 SI 1990 No 1431 (the Food Hygiene Amendment Regulations).
7 *Welton* v *North Cornwall District Council*, n 3 above, 573, *per* Rose LJ.
8 *ibid*.
9 *ibid*.
10 *ibid*, 581, *per* Rose LJ.
11 Judge Thompson found that 90 per cent of the work carried out by the plaintiffs with the aim of complying with the law was unnecessary This point is noted at *ibid* 574, *per* Rose LJ.
12 The four conditions described in the text are set out in *Caparo Industries plc* v *Dickman* [1990] 2 AC 605, 638, *per* Lord Oliver.
13 n 3 above 578 and 580, *per* Rose LJ.
14 *ibid*, 574, *per* Rose LJ
15 *ibid*, 585, *per* Ward LJ.
16 See *ibid*, 581, where Rose LJ identifies Mr Evans as having made threats of 'closure and close supervision'. See also 582 and 586, *per* Ward LJ and 587, *per* Judge LJ.

© The Modern Law Review Limited 1999

Court saw the fact that Mr Evans had made threats as strengthening the argument in favour of imposing liability. In this regard, Judge LJ stated that the threatening character of the statements made by Mr Evans made it '*more important* that they should be included within the ambit of [the *Hedley Byrne* principle]'.[17] Further, the facts of the case were identified by Rose, Ward, and Judge LJJ as disclosing an 'assumption of responsibility' on Mr Evans' part.[18] Hence, he had been under a duty to exercise reasonable care when telling the plaintiffs what they had to do in order to comply with the law. Such care had not been taken. This being so, the Court of Appeal concluded that the plaintiffs could recover for the unnecessary expenses they had incurred.

By way of support for their decision, Rose, Ward and Judge LJJ emphasised the *Hedley Byrne* principle's breadth. While recognising that it is typically associated with the giving of information and/or advice, they identified it as comprehending the exercise of a wide range of 'special skills'. Moreover, they recognised that such skills can be exercised in the course of providing a variety of services. The three judges found support for this view in the following passage from Lord Goff's speech in *Henderson* v *Merrett Syndicates Ltd*:

> [T]he facts of *Hedley Byrne* itself ... show that the concept of a 'special skill' must be understood *broadly* .... [T]hough *Hedley Byrne* was concerned with the provision of information and advice, the example given by Lord Devlin of the relationship between solicitor and client ... show[s] that *the principle extends beyond the provision of advice and information to include the performance of other services.*[19]

Before upholding the trial judge's decision, the Court of Appeal satisfied itself that the imposition of liability would not be inconsistent with the efforts of local authorities to carry on their environmental health-related statutory functions.[20] In this connection, the Court identified two reasons as to why no such inconsistency would arise. First, '[Mr Evans] was acting outwith the powers of the [relevant] Acts and also outside the informal enforcement practice of the [defendant Council]'.[21] Secondly, the liability rule that they were establishing was addressed only to conduct such as that of Evans. Hence, it could reasonably be expected not to constrain councils when carrying on their statutory functions. Perhaps the strongest argument in support of this conclusion can be found in Rose LJ's judgment. He identifies three kinds of conduct that might prompt the imposition of liability under *Hedley Byrne*. They are:

(i) 'conduct [that is] specifically directed to statutory enforcement, such as the institution of proceedings before justices, the service of an improvement notice and the obtaining of a closure order';[22]
(ii) conduct that takes the form of 'offering ... an advisory service';[23]

17 *ibid*, 587 (emphasis added).
18 See *ibid*, 580, *per* Rose LJ, 582, *per* Ward LJ, and 587, *per* Judge LJ. See also 582–583, where Ward LJ draws an analogy between the parties' relationship in *Welton* and that of the parties in *Henderson* v *Merrett Syndicates Ltd*. In the latter case, the House of Lords justified the imposition of liability by reference to the idea of an assumption of responsibility.
19 *ibid*, 579–580, *per* Rose LJ (quoting from *Henderson* v *Merrett Syndicates Ltd*, n 2 above, 180, *per* Lord Goff).
20 *ibid*, 584–586, *per* Ward LJ. (In seeking to ensure that the imposition of a duty of care would not be inconsistent with the defendant local authority's statutory functions, the Court followed the lead given in *X (Minors)* v *Bedfordshire County Council* [1995] 2 AC 633, 739, *per* Lord Browne-Wilkinson.)
21 *ibid*, 584, *per* Ward LJ.
22 *ibid*, 580.
23 *ibid*, 581.

© The Modern Law Review Limited 1999

(iii) conduct that involves 'the imposition ..., outwith the legislation, of detailed requirements enforced by threats of closure and close supervision'.[24]

Rose LJ identifies 'the conduct which is at the heart of [*Welton*]' as falling within the third of the above categories.[25] This being so, the rule of liability established by the Court can be seen to be very narrowly focused. Moreover, this rule comprehends forms of activity that cannot be regarded as serving to advance the purposes of the legislation under which Mr Evans was purporting to act. Hence, a basis exists for concluding that the liability rule established in *Welton* should not, other things being equal, prove inconsistent with the efforts of local authorities to carry out their statutory functions.

In the light of the points noted above, the three Lords Justices concluded that the rule of liability they were specifying fell '*well within the Hedley Byrne* principle'.[26] This is a phrase that throws considerable light on the Court's understanding of the adjudicative task in which it was engaged. To talk of a fact-situation falling within a principle is to employ a *spatial* metaphor. It is to conceive of the *Hedley Byrne* principle as what might be termed a field of interpretative possibility: ie a normative 'space' within which a range of (particular) liability rules can legitimately be established. It is unsurprising that we should find the Court thinking in this way. Those responsible for the enunciation of the *Hedley Byrne* principle used highly abstract language. To take an obvious example, Lord Reid stated in *Hedley Byrne* that liability can arise in circumstances where the parties stand in a 'special relationship'.[27] Language of this sort is underdeterminate and, hence, open to a range of plausible interpretations.[28]

In stating that the liability rule established in *Welton* falls squarely within the *Hedley Byrne* principle, the Court of Appeal conveys the impression that it is making an uncontroversial decision. Much the same impression is created by the Court's efforts to specify the *ratio* of its decision as narrowly as possible. Relevant, in this regard, are the distinctions drawn by Rose LJ between the three types of situation in which *Hedley Byrne* might be applied to the activities of environmental health officers.[29] As we have already noted, Rose LJ states that the liability rule fashioned in *Welton* only applies to the last, and narrowest, of the three types of situation he describes. Hence, he can be regarded as having adopted an approach to his task that provides an example of what the American commentator Cass Sunstein has termed 'decisional minimalism'. This is an approach to adjudication that involves judges in (a) 'decid[ing] no more than they have to decide' and (b) in seeking to leave questions that are not presented to them 'open'.[30] Moreover, Sunstein notes that this approach to adjudication is much favoured by Anglo-American judges. This is because it serves, *inter alia*, to limit the scope of the inherently controversial task of judicial law-making.[31] Hence, it is unsurprising

24 *ibid*, 581.

25 *Ibid.*

26 *Ibid*, (emphasis added), *per* Rose LJ. See also 582 (emphasis added), *per* Ward LJ (stating that '[t]he speech of Lord Morris of Borth-y-Gest in *Hedley Byrne* is *directly applicable* to the facts of [*Welton*]') and 589, *per* Judge LJ (where it is stated that 'the present case does not involve any creation or extension of principle').

27 *Hedley Byrne & Co Ltd* v *Heller & Partners Ltd*, *supra*, n 1, 486.

28 See C. Kutz, 'Just Disagreement: Indeterminacy and Rationality in the Rule of Law' (1994) *Yale LJ* 997, 1002, where it is noted that 'multiple conclusions' can be drawn from underdeterminate legal propositions

29 See ns 22–24 above (and associated text).

30 See C. Sunstein, 'Foreword: Leaving Things Undecided' (1996) 110 *Harv LR* 6, 6–7.

31 *ibid*, 7–8, 10 and 26.

© The Modern Law Review Limited 1999

that we find Rose LJ and his colleagues making strenuous efforts to present their decision in the terms noted above. There is, however, reason to conclude that the decision in *Welton* is much bolder than the three Lords Justices concede. Support for this view can be found in the fact that *Hedley Byrne* has now been identified as comprehending threats. This point is developed below in the course of a discussion of the *Hedley Byrne* principle's scope.

The scope of the *Hedley Byrne* principle

The *Hedley Byrne* principle is, as noted above, couched in underdeterminate language. Hence, it is unsurprising that it has been applied in both two-party and three-party situations and in both business and non-business contexts.[32] While the *Hedley Byrne* principle has been applied in a wide range of circumstances, common to these situations is the following feature. Defendants have not acted in ways that have constituted a curtailment of the autonomy of the plaintiffs upon whom they have wrongfully inflicted harm.[33] This is, for example, true of those defendants who have given plaintiffs reasons for action, in the form of information and/or advice, that the latter have voluntarily acted upon to their detriment.[34]

Mention of the activity of giving reasons for action brings us to a distinction that is of central importance to this discussion. While, in *Welton*, Mr Evans gave the plaintiffs reasons for action, those reasons can be regarded as being, in moral terms, qualitatively different from those provided in earlier *Hedley Byrne* cases. The beginnings of an explanation as to why this view can be taken can be found in the fact that, in *Welton*, the *Hedley Byrne* principle was applied to a coercive utterance. This utterance took the form of a conditional threat. Such a threat is issued in circumstances where one person (A) says to another (B): 'unless you do X (X being a specified course of conduct), I will do Y (Y being something that can reasonably be regarded as harmful, potentially harmful, or, in some significant sense, unpleasant)'.[35] In *Welton*, Mr Evans made a number of statements that can be reduced to the following conditional threat: 'unless you [ie the plaintiffs] meet the requirements stipulated, then your business will be closed down'. In light of this last and the foregoing points, the statements made by Evans can be regarded as inducements to act so strong as to constitute a curtailment of the plaintiffs' autonomy. It is this point that supports the view that the statements in *Welton* are, in moral terms, qualitatively different from those in earlier cases in which *Hedley Byrne* has been applied.[36]

32 For discussion of the two- and three-party situations in which the *Hedley Byrne* principle has been applied, see R. Epstein, *Simple Rules for a Complex World* (London: Harvard University Press, 1995), 198–206. For discussion of the use made of *Hedley Byrne* in both business and non-business contexts, see D. Howarth, *Textbook on Tort* (London: Butterworths, 1995), 275–279.

33 On 'autonomy', see J. Raz, *The Morality of Freedom* (Oxford: Clarendon Press, 1986) 155, where it is defined as the substantial authorship of one's life.

34 For further discussion of reasons for action, see H.L.A. Hart and A.M. Honoré, *Causation In The Law* (Oxford: Clarendon Press, 2nd edn, 1985) 2 and 51–59.

35 See K. Greenawalt, *Speech, Crime, And The Uses of Language* (Oxford: Oxford University Press, 1989), 91 (on conditional threats). See also H.L.A. Hart, *The Concept of Law* (Oxford: Clarendon Press, 2nd ed, 1994) 19 (on the coercive character of threats to do 'something that a normal person would regard as harmful or unpleasant').

36 A further moral distinction (closely related to the one discussed in the text) can be drawn between *Welton* and *Hedley Byrne*. All acts of coercion (including, of course, those in *Welton*) can be regarded as intentional (see Raz, n 33 above, 148–149 and 378). Assuming that this view is correct, it buttresses the argument that the conduct at issue in *Welton* should have been identified as falling outside *Hedley Byrne*, which has, of course, to do with *careless* activity. Against this point must, however, be set the

© The Modern Law Review Limited 1999

The foregoing analysis, if correct, provides a basis for suggesting that the Court of Appeal erred in claiming that it could derive support for its decision from Lord Goff's analysis in the *Henderson* case. In *Henderson*, Lord Goff (as already noted) associated the *Hedley Byrne* principle with the provision of services. The term 'service' is, of course, quite strongly associated with, *inter alia*, the giving of information and/or advice.[37] It is not, however, associated with the specification of 'requirements' that *must* be met in order to avoid subjection to legal sanctions.[38] Pertinent to the point of contrast here being drawn is the concept of reliance (which the Court of Appeal invoked as a ground for its decision in *Welton*).[39] 'Reliance' (as placed by plaintiffs on the statements of defendants) is, of course, a concept with which *Hedley Byrne* is strongly, if not invariably, associated.[40] The *Hedley Byrne* principle's strong association with this concept can be explained (at least in part) by reference to the services with which it is centrally concerned: *viz*, the provision of information and/or advice. Those who avail themselves of such services can be presumed to be seeking material on which they may voluntarily base their conduct. Further, to the extent that the conduct of an advisee is based on such material, a relationship of reliance can be said to exist. But talk of reliance surely becomes otiose in circumstances where one person presents another (or others) with 'requirements' that must be met in order to avoid the imposition of legal sanctions. This being so, a basis exists for doubting that Mr Evans was engaging in a mode of activity that can aptly be characterised as the provision of a service. This is, of course, a point that lends further support to the view that his conduct ought not to have been identified as falling within *Hedley Byrne*.

At this point, it may be objected that the Court of Appeal's decision in *Welton* provides an example of the legitimate judicial modification of a malleable legal principle. This is an objection that can be developed by employing a form of reasoning that involves the pursuit of what the philosopher John Rawls terms reflective equilibrium.

Reflective equilibrium and the *Hedley Byrne* principle

Reflective equilibrium is pursued in circumstances where both the principles that generate a conclusion and the conclusion itself are revised in the light of each other. The aim of this process is to bring abstract principles and the concrete conclusions that they have been taken as supporting into a reasonable degree of harmony.[41] This form of reasoning provides the basis of a three-step argument, the burden of which is that the facts of the *Welton* case can plausibly be brought within the *Hedley Byrne* principle. The first step in this argument is to concede that a

consideration that the conduct at stake in *Welton* is a *composite* of the careless (the thirteen inaccurate statements) and the coercive (the making of a conditional threat). Hence, in this discussion, reliance has been placed on the distinction between conduct that curtails autonomy and that which does not. This latter distinction provides an adequate basis upon which to argue that *Welton* should not have been categorised as a *Hedley Byrne* case.

37 Support for the view set out in the text can be found in *The New Shorter Oxford English Dictionary*, vol II, edited by L. Brown (Oxford: Clarendon Press, 1993) 2789.

38 *ibid.*

39 See n 14 above, where the Court of Appeal's invocation of the concept of reliance is noted.

40 See C Barker, 'Unreliable Assumptions in the Modern Law of Negligence' (1993) 109 LQR 461, 481, for discussion of one sub-set of negligent misstatement cases (the 'wills cases') in which reliance has not been placed by plaintiffs on the statements of defendants. For an example of such a case, see *White* v *Jones* [1995] 2 AC 207.

41 See J. Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1971) 20 *et seq.*

© The Modern Law Review Limited 1999

tension exists between *Welton* and *Hedley Byrne* as understood prior to the Court of Appeal's decision. The argument's second step is to move from the level of concrete conclusion (the decision in *Welton*) to abstract principle (*Hedley Byrne*) and to revise the latter in the light of the former. This can be done by identifying the *Hedley Byrne* principle as comprehending not only the provision of information, advice, and other services but, also, the issuing of threats (of the sort in *Welton*). Having set out this revised understanding of the principle, the argument's third step is to return to *Welton* and identify the Court of Appeal as having pointed up *Hedley Byrne's* commodiousness. Having worked through this argument, a basis can be said to exist for claiming that no significant tension exists between the decision reached in *Welton* and the (now better understood) *Hedley Byrne* principle. This argument (the second step of which exhibits considerable interpretative fiat) is, however, open to two objections. These objections are set out below. So too is a suggestion as to how the claim in *Welton* might have been better accommodated within the law of tort.

Law, legal language and novel wrongs

The first of the objections to the argument outlined above has to do with the nature of law as an institution. To establish a system of law is to seek to control or regulate practical affairs *over time*.[42] It is to establish a conservative institution in which the past (in the form of, for example, already articulated liability rules) is supposed to control or regulate distinct sets of (present) circumstances.[43] In order for this to happen, established norms must be respected. They must be implemented either as specified or (to take a broader view) in accord with their informing purposes.[44] In light of this last and the foregoing points, the reflective equilibrium-based argument canvassed above can be seen to sit uneasily with the institution of law. This is because it is an argument in which too little regard is paid to the principle of liability established in *Hedley Byrne*. Two points provide support for this view. First, the statements in *Welton* were (as already noted) coercive. Hence, they can be regarded as lying outside the *Hedley Byrne* principle as understood since the time of its enunciation. Secondly, the Court of Appeal's decision is identified as providing a basis upon which to modify the very norm that is said to authorise that decision.[45] Modification of this sort involves an inversion of the way in which law is, on the above account, supposed to operate. This is because *present circumstances* (as embraced within an expanded *Hedley Byrne* principle) are identified not as an object of legal regulation but, rather, as a driver of what is presented as legal change.

The second objection to the reflective equilibrium-based argument canvassed above concerns legal language. The language in which duty-imposing norms are

42 Support for the view advanced in the text can be found in L. Fuller, *The Morality of Law* (New Haven: Yale University Press, revd edn, 1969) 79 *et seq*.

43 *ibid*, 53 (emphasis added), where it is stated that 'the proper movement of law is *forward in time*'. Support for the application of the adjective 'conservative' to the institution of law can be found in A. Giddens, *Beyond Left and Right: the Future of Radical Politics* (Cambridge: Polity Press, 1994). See 45, where the view that the past should control the present is characterised as conservative. See also 24–25, where conservatism is identified as being 'bound up with ... practices' that promote the continuity of social institutions.

44 See J. Waldron, 'Kant's Legal Positivism' (1996) 109 *Harv LR* 1536, 1538–1539.

45 cf S. Fish, *There's No Such Thing As Free Speech And It's A Good Thing Too* (Oxford: Oxford University Press, 1994) ch 11 (for discussion of processes of adjudication in which legal norms are modified in the way described in the text).

© The Modern Law Review Limited 1999

couched works by uniting the general (eg an authoritative norm) and the particular (eg the sets of circumstances identified as being comprehended by that norm).[46] It does this by establishing points of relevant similarity between a number of fact-situations in which duties are identified as existing. The *Hedley Byrne* principle, for example, establishes such points of similarity between a range of circumstances in which services (most obviously, the provision of information and/or advice) are proffered.[47] There are, however, limits to the extent to which a norm can embrace a range of circumstances and yet retain its integrity as a distinct and, as such, readily intelligible guide to action.[48] Support for this claim can be found in the Court of Appeal's decision in *Welton*. Prior to this decision, the *Hedley Byrne* principle had, as noted above, been associated with the provision of services. It had not been associated with the issuing of threats. Moreover, the statements made in *Welton* can, as already noted, be regarded as qualitatively different from those made in earlier cases. This is, of course, because of the autonomy-curtailing impact of threats. This being so, a basis exists for suggesting that, to embrace the facts of *Welton* within *Hedley Byrne* is to seek to present as essentially the same two modes of interaction that are irreducibly different. This is a course of action that has the effect of making the *Hedley Byrne* principle a less distinct and, as such, a less intelligible, norm.[49]

The difficulties described above might have been avoided had the plaintiffs presented their claim as a hitherto unrecognised form of wrong, sufficiently serious to justify the specification of a new head of tort liability.[50] Three points support this view. First, the statements at stake in *Welton* can, on the above analysis, be regarded as falling outside the *Hedley Byrne* principle. Secondly, the harm inflicted on the plaintiffs by Mr Evans does not appear to be comprehended by any other existing head of liability. A claim could not have been based on the tort of misfeasance in public office. This is because Mr Evans was not actuated by malice.[51] Likewise, an action for interference with trade could not have been mounted, since Evans did not intend to injure the plaintiffs.[52] Thirdly, a basis exists for arguing that the plaintiffs were treated in a manner that could have prompted the specification of a new head of liability. New heads of liability have typically been established in tort in circumstances where the following three conditions have been satisfied. First, the plaintiff has a significant interest in that upon which harm has been inflicted.[53] Secondly, harm has been suffered in the context of a wrongful

46 For a particularly clear formulation of the point made in the text, see T. Eagleton, *William Shakespeare* (Oxford: Blackwell, 1986) 41.

47 See B. Feldthusen, *Economic Negligence: the Recovery of Pure Economic Loss* (Scarborough, Ontario Thomson, 3rd ed, 1994) ch 2. See esp 63.

48 See T. Eagleton, *The Illusions of Postmodernism* (Oxford: Blackwell, 1996) 37, where support can be found for the proposition that, as the range of reference of a verbal formulation is expanded, its intelligibility can be reduced

49 See P. Cane, *The Anatomy of Tort Law* (Oxford: Hart Publishing, 1997) 213, for discussion of the relationship between the internal consistency of legal norms and their intelligibility.

50 Another way of putting the point made in the text would be to say that *Welton* should have been argued as an action on the case. For discussion of actions on the case, see J.H. Baker, *An Introduction To English Legal History* (London: Butterworths, 3rd ed, 1990) 73–75.

51 See P. Cane, n 49 above, 54–55. (While Mr Evans can be said to have acted intentionally insofar as he acted coercively (see n 36 above), this would not suffice to meet the (narrow) 'malice' requirement of the tort discussed in the text)

52 *ibid*, 33.

53 See L. Dolding and R. Mullender, 'Tort Law, Incrementalism, And The House of Lords' (1996) 47 NILQ 12, 12. (On the analysis set out in this piece, plaintiffs have a significant interest in those goods that constitute necessary, if not sufficient, conditions of their leading autonomous lives. Examples of such goods are provided by physical and financial security.)

 © The Modern Law Review Limited 1999

transaction with the defendant.[54] Thirdly, the benefits of a finding of liability could not, other things being equal, reasonably be expected to be outweighed by countervailing costs.[55] The facts of *Welton* provide a basis upon which to argue that each of the above conditions could be satisfied. First, the plaintiffs could be said to have a significant interest in not suffering the financial harm consequent upon acting in the manner specified by Mr Evans.[56] Secondly, the harm suffered by the plaintiffs can reasonably be regarded as having occurred in the course of a wrongful transaction with Mr Evans. Support for this suggestion can be found in the latter's having coerced them, without legal warrant, into embarking upon a financially damaging course of action. Thirdly, the positive value of a liability rule could not, *ceteris paribus*, reasonably be expected to be outweighed by countervailing costs. This is because (as the Court of Appeal demonstrates in *Welton*) a liability rule narrowly tailored to the facts of the case could be established. Such a rule could reasonably be expected not to stymie the legitimate activities of local authorities.

At this point, it must be conceded that the above proposal is open to the criticism that it sits uneasily with the ideal of the rule of law. This is because acting on it would involve the retrospective articulation of a new head of liability. This objection cannot be gainsaid. There is, however, one consideration that provides a basis upon which to suggest that the enunciation of a new head of liability would have been preferable to the course of action taken in *Welton*. To have acted on the above proposal would have been to make transparent the strongly innovative character of the step being taken.

Conclusions

On the above analysis, the Court of Appeal has applied the *Hedley Byrne* principle to a set of facts that should not be regarded as being comprehended by it. This is because it has identified as falling within the principle a coercive form of utterance that is qualitatively different from those statements to which it has hitherto been applied. While, for this reason, the Court's decision is regrettable, it is not altogether surprising. In recent years, the *Hedley Byrne* principle has been repeatedly described as broad in scope.[57] Moreover, *Hedley Byrne* has, on at least

54 See *ibid*, 13–14, where it is argued that significant interests are typically protected by tort law in circumstances where a defendant has inflicted harm on a plaintiff in the course of a wrongful transaction. (It is also argued in this passage that a wrongful transaction can be said to occur where a defendant's conduct bespeaks a lack of proper regard for the interest on which harm is inflicted. Hence, the relevant conduct can be categorised as wrongful and not merely as harmful.)

55 See *ibid*, 14–15.

56 It might also be suggested that persons such as the plaintiffs in *Welton* have a significant interest in not being coerced without justification. This is an interest that could be protected by a tort requiring a showing of intention, which (on the analysis set out at n 36 above) is a feature of coercive activity. Such a requirement would, of course, limit the scope of a new head of liability.

57 The point made in the text is not intended to convey the impression that the Court of Appeal's decision in *Welton* was the inevitable upshot of recent judicial emphasis on *Hedley Byrne's* breadth. Relevant in this regard is the case of *Harris* v *Evans and Another* [1998] 3 All ER 522 (CA). *Harris* concerned a health and safety officer who gave advice to a local authority. The advice given prompted the local authority to impose requirements (improvement and prohibition notices) on the plaintiff. Consequently, the plaintiff's business sustained economic loss. The Court of Appeal held that *Hedley Byrne* was inapplicable to the advice given by the officer. At 538, Sir Richard Scott V-C justified the Court's decision by reference, *inter alia*, to the 'seriously detrimental' effects that could reasonably be expected to flow from the specification of a liability rule. In this connection, he stated that the rule argued for by the plaintiff could be expected to compromise health and safety officers' efforts to promote the public interest in workplace-

© The Modern Law Review Limited 1999

one occasion, been identified by a Law Lord as embracing a set of facts not obviously within its scope. In *Murphy* v *Brentwood District Council*, Lord Keith identified the *Hedley Byrne* principle as embracing the facts of *Junior Books* v *Veitchi Corp*.[58] This was, of course, a case in which a sub-contractor was held liable for laying defective (but not dangerous) flooring in the plaintiff's business premises. Lord Keith's characterisation of *Junior Books* has, since the time of its articulation, been a source of perplexity in the legal community.[59] This is because the facts of the case are so far removed from fact-situations of the sort with which *Hedley Byrne* is associated. So great are these differences that Lord Keith's characterisation of *Junior Books* can be regarded as involving an abuse of the law of negligence.[60] One way of capturing what such abuse consists in is to see it as an indefensible departure from the field of interpretative possibility within which Lord Keith was claiming to base his analysis. Central to the analysis advanced in this piece is, of course, the claim that the facts of *Welton* lie outside the same field. If this claim is correct, then the Court of Appeal would have been better advised to give expression to its impulse to find for the plaintiffs by establishing a new head of liability.

related safety. He concluded that such a state of affairs would be unacceptable since it could not be reconciled with the objectives of the Health and Safety at Work Act 1974. Moreover, Scott V-C distinguished *Welton*, having confessed to 'some difficulty' with it (538) He saw it as distinguishable since, on his analysis, no consideration was given by the Court of Appeal to the relevant 'statutory framework'. This criticism seems misplaced. In *Welton*, the Court emphasised the importance of ensuring that the imposition of a duty of care would not compromise the efforts of councils to promote the public interest in hygienic food preparation. See, in this connection, ns 20–24 above (and associated text).

58 Lord Keith's comment can be found at [1991] AC 398, 466. *Junior Books* v *Veitchi Corp* is reported at [1983] AC 520.

59 D. Howarth, *Textbook on Tort*, n 32 above, 293.

60 Support for the view set out in the text can be found in N.J. McBride and A. Hughes, '*Hedley Byrne* in the House of Lords: an Interpretation' (1995) 15 LS 376, n 44.

© The Modern Law Review Limited 1999

Pregnancy Discrimination: Equality, Protection or Reconciliation?

Michael Wynn*

Dismissal for pregnancy-related sickness has been justified on financial and administrative grounds for many years. Employers maintained that they should not be burdened with the financial consequences of pregnancy and maternity where this entailed long term incapacity even where subsequent dismissals infringed the principle of equal treatment. Ever since the twin decisions of *Dekker*[1] and *Hertz*,[2] the European Court of Justice has struggled to find an acceptable balance between the equality principle and the demands of a male oriented labour market. *Hertz* tipped the balance in favour of the latter by ruling that dismissal for absence after maternity leave was not discriminatory even where the illness originated in pregnancy. The Pregnant Workers Directive[3] which introduced special protection for pregnant women throughout the pregnancy and maternity period, left the scope of *Hertz* unclear. The Court has now clarified this vexed area of law in *Brown* v *Rentokil Ltd*[4] by ruling that it is direct discrimination to dismiss a woman at any time during her pregnancy for absences caused by pregnancy-related illnesses. The Court has not, however, chosen to depart from the limitations imposed in *Hertz* and followed in *Larsson*.[5] *Brown* highlights the limitations of the equality concept in dealing with female difference in the workplace, as positive discrimination in favour of motherhood is circumscribed by artificial time limits which are left to be defined under the principle of subsidiarity. This is another pragmatic decision which recognises that, for cost reasons, domestic and work roles must conform with male work patterns outside a narrowly drawn maternity period.

The facts

Mary Brown was employed as a driver for Rentokil, servicing Sanitact units in retail outlets. She regarded this job as heavy work. When she became pregnant in August 1990, she informed Rentokil. She was absent from work from the middle of August onwards and submitted certificates for symptoms described as 'bleeding in pregnancy' and 'pregnant backache'. She never worked again. Rentokil operated a working rule under the contracts of employment of all its employees whereby any employee who was absent for sickness for more than twenty six weeks

* Faculty of Management and Business, University College Northampton
Thanks to James Hanlon and the anonymous referee for helpful comments on an earlier draft.

1 Case C-177/88, *Dekker* v *Stichting Vormingscentrum Voor Jonge Volwassenen (VJV-Centrum) Plus* [1990] ECR I-3941.

2 Case C-179/88, *Handels-Og Kontorfunktionaerernes Forbund I Danmark* v *Dansk Arbejdsgiverforening* [1990] ECR I-3979.

3 Directive 92/85/EC adopted on 19 October 1992, on the introduction of measures to encourage improvements in the safety and health at work of pregnant workers and workers who have recently given birth or are breastfeeding (OJ 1992 L 348, 1).

4 Case C-394/96 [1998] 2 CMLR 1049.

5 *Handels-Og Kontorfunktionaerernes Forbund i Danmark* v *Dansk Arbedjdsgiverforening* Case C-400/95 [1997] ECR I-2757.

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

continuously would be dismissed. They informed Mrs Brown in November 1990 that if she did not return to work before the end of the twenty six week period after an independent medical examination, her employment would be terminated in accordance with the working rule. Mrs Brown was dismissed, while still pregnant, in February 1991.

The dispute before the national courts

The litigation arose as a result of the fact that Mrs Brown was not entitled to any statutory maternity rights as she did not have the requisite two year period of continuous employment, thus she had no right to take absence from work for maternity leave or to return to work in the twenty nine weeks following delivery. By the same token, she had no claim for unfair dismissal for a reason connected with pregnancy under section 60 of the Employment Protection (Consolidation) Act 1978.[6] Mrs Brown brought a claim for sex discrimination, alleging that her dismissal was based on the fact of her pregnancy.

The Scottish industrial tribunal dismissed her claim under the Sex Discrimination Act 1975 on the grounds that her dismissal was not automatically protected as being discriminatory where the pregnancy-related illness fell outside the statutory maternity provisions. The Employment Appeal Tribunal dismissed Mrs Brown's appeal, considering itself bound by the decision of the Court of Appeal in *Webb* v *EMO Air Cargo UK Ltd.* The Court of Session held that the facts of *Webb* were different and declared that there was a distinction under the Equal Treatment Directive[7] between dismissal for pregnancy itself and dismissal for illness caused by pregnancy which the European Court of Justice had recognised in *Hertz.*

Mrs Brown appealed to the House of Lords, which referred the following question to the European Court of Justice: is it contrary to the principle of non-discrimination under Directive 76/207/EEC for an employer to dismiss a pregnant worker for incapacity for work caused by pregnancy when the period of absence from work has exceeded the maximum sick leave stipulated in the contract of employment? There were two parts to this question, first, was it permissible under Community law to dismiss a pregnant employee at any time during the pregnancy for such absences, and secondly, did the contractual provision make any difference to the outcome? The Court of Justice ruled that the protection against sex discrimination under Directive 76/207/EEC extended to all illnesses attributable to pregnancy which arose during the time of pregnancy. The Court also decided that the contractual term made no difference to the outcome.

The Advocate General's opinion

Advocate General Colomer applied a simple equal treatment formula to the questions under consideration.[8] The principle of non-discrimination under Article 5 of Directive 76/207/EEC proscribes the application of the same rule to different

6 An employee in her position would now be protected under s 99 Employment Rights Act 1996 implementing Directive 92/85/EC.

7 Directive 76/207/EEC of 9 February 1976 on the implementation of the principle of equal treatment for men and women as regards access to employment, vocational training and promotion, and working conditions (OJ 1976 L 39, 40).

8 He stated that the time limit for implementation of Directive 92/85/EC had not yet expired when Mrs Brown was dismissed at the beginning of 1991, thus this Directive was not directly binding.

© The Modern Law Review Limited 1999

situations or different rules to comparable situations. As pregnancy only affects women, any resulting incapacity requiring absences from work occurring within the pregnancy could not be compared to male absences for sickness.

The particular problem in *Brown*, however, was to find a satisfactory distinction between sicknesses which could be attributed to pregnancy and those which fell outside this category. Both *Hertz* and *Larsson* which purported to follow *Hertz*, had suggested that illnesses occurring outside the period of maternity leave were to be treated as ordinary sicknesses and outside the special protection afforded to pregnancy which did not require comparison with the treatment of a man, but within the scope of comparison for equal treatment, even though the illness originated in pregnancy. Thus, if the sickness absence would have led to the dismissal of a male worker under the same conditions, there would be no direct discrimination on grounds of sex.[9] The Advocate General was prepared to accept that 'high-risk pregnancies' might incapacitate a woman for the whole period of pregnancy. Many pregnancy-related conditions were not akin to standard illnesses as they were conditions requiring strict medical supervision and necessitating repeated absences from work but nevertheless, these conditions were of a temporary nature.[10]

The main distinction which brought a pregnant woman within the principle of non-discrimination was the fact that pregnancy was not only female-specific, but in most cases, involved a finite period of incapacity. Thus a rule could be formulated according to the categorisation made in *Hertz* between the 'normal risks of pregnancy'[11] which occurred in the period of pregnancy and confinement and conditions which fell outside this period. Illness which arose from pregnancy could then be classified by a simple time reference: if it manifested during the pregnancy and maternity leave period it was protected, if it arose later, it was not. Thus:

> if one wishes to draw a distinction between the normal risks of pregnancy and confinement and medical conditions which do not reflect the ordinary risks of pregnancy, it is necessary to take a chronological approach: the former will, necessarily, be those which arise while the woman is in one of those situations, that is to say whilst she is pregnant or on maternity leave, whereas the latter will be all those which arise after maternity leave, even though their cause may be traceable back to the pregnancy or confinement.[12]

The judgment of the European Court of Justice

In concluding that dismissal during pregnancy for incapacity to carry out any contractual duty was contrary to Directive 76/207/EEC, the Court of Justice agreed with the Advocate General that the condition of pregnancy requiring special protection naturally encompassed those disorders arising from pregnancy itself. These risks deserved protection as an inherent risk of pregnancy.[13]

The entire period of pregnancy was subject to the same risks and a dismissal for any such condition was directly discriminatory[14] as it was based on the fact of

9 The problem was particularly acute in *Larsson* as most of the sickness had occurred within the pregnancy period itself but the employer had not dismissed the worker until after her statutory maternity leave had expired.

10 *Brown* v *Rentokil*, n 4 above, Opinion of AG Colomer, 5 February 1998, para 59.

11 *Hertz*, n 2 above, Opinion of AG Darmon, at para 48.

12 *Brown* v *Rentokil*, n 4 above, Opinion of AG Colomer, para 65.

13 *Ibid*, Judgment of 30 June 1998, para 22.

14 It could equally be argued that such treatment is indirect discrimination on the grounds of adverse impact on women as a group. Any justification on grounds of business necessity could then be made more explicit. See R Wintemute, 'When is Pregnancy Discrimination Indirect Discrimination?' (1998) ILJ 23.

© The Modern Law Review Limited 1999

pregnancy. As a result, 'the principle of non-discrimination required similar protection throughout the period of pregnancy'.[15]

Having decided that special pregnancy protection extended up to the end of maternity leave and no further, it was logical that conditions arising after the end of maternity leave should not be protected as long as the woman's absence was treated in the same way as a man's absence for incapacity; thus the Court confirmed its decision in *Hertz*. However, the Court doubted the correctness of its previous decision in *Larsson*. Hertz had resumed work for a year after maternity leave before she became ill again whereas Larsson's sick leave occurred mainly during the pregnancy itself. If the whole pregnancy period was protected, it was not lawful to dismiss for pregnancy sickness which had occurred in that period:

> her absence not only during the maternity leave but also during the period extending from the start of her pregnancy to the start of her maternity leave cannot be taken into account for computation of the period justifying her dismissal under national law.[16]

On the second question as to the effect of the contractual absence clause, the Court emphasised that equality must take precedence over freedom of contract. As pregnancy was a female condition, the application of a contractual term enabling dismissal of both male and female workers for long term sickness absence was clearly discriminatory.

Clarification and confluence

Brown has clarified many of the questions relating to the legal protection afforded to women dismissed for pregnancy-associated sickness. The result of the case is that the entire period of pregnancy and maternity leave is now a special protected period during which both Directives 76/207/EEC and 92/85/EC prohibit pregnancy-related dismissals on grounds of equality and health and safety respectively. The ambit of the two Directives has been clarified: the total prohibition on dismissal in the pregnancy and maternity leave periods in Article 10 of Directive 92/85/EC means that the equal treatment protection under the earlier Directive will only be relevant in a *Larsson* type situation where a woman has been dismissed after maternity leave for an earlier pregnancy-related illness. The ambiguity in *Hertz* as to the effect of a post-maternity dismissal has been resolved. Illnesses caused by pregnancy are protected unless they manifest after the maternity leave has ended. After this point, the equality formula allows business exigency to determine liability under the 'sick man' rule. The general approach to pregnancy illness is to take a chronological perspective to the onset of illness and classify protected and non-protected conditions according to the 'normal risks of pregnancy' occurring in the protected period.

Nevertheless, *Brown* presents further difficulties. The decision represents the confluence of two separate approaches to pregnancy discrimination. Whereas Directive 76/207/EEC guarantees equal treatment in equal situations, Directive 92/85/EC permits special treatment of the pregnancy condition on grounds of health and safety.[17] The central weakness of the equality approach to issues of discrimination at work has been highlighted in cases of pregnancy dismissal where the

15 *Brown* v *Rentokil*, n 4 above, Judgment, para 24

16 *ibid*, para 27.

17 Directive 92/85/EC, in effect, permits positive discrimination in favour of pregnant women provided that a worker can satisfy the conditions imposed by the Directive.

© The Modern Law Review Limited 1999

search for a male comparator has led to reliance on a male norm and resulted in a comparison with a sick male.[18] *Brown* has not entirely eliminated the sick-male approach in that it is still permissible to use such a comparison where the illness arises after the end of maternity leave. The Advocate General stated in *Brown* that conditions arising after maternity were not associated with the 'ordinary risks of pregnancy' and should be treated in the same way as ordinary sickness.[19] A more robust approach which favoured a female norm would recognise female difference in the workplace and attempt to eliminate any disadvantage arising from this difference.[20] Although the Court stated in *Brown* that both Directive 76/207/EEC and Directive 92/85/EC recognise the legitimacy of protecting the nurturing function of motherhood as well as pregnancy itself,[21] in practice, the wider protection was severely limited. The advantage of a health and safety perspective towards pregnancy discrimination under Directive 92/85/EC is that a 'risks' approach to pregnancy avoids explicit comparison with male norms, but the danger is that a focus on risks peculiar to females will be constructed on organisational patterns which do not necessarily value parenthood itself. More recent EC initiatives, for example, the Parental Leave Directive,[22] are beginning to address the wider issues of reconciling work with domestic responsibilities.[23]

The neat solution arrived at in *Brown* will obviate much further litigation on the extent of protection provided for pregnancy illnesses, but residual problems will still arise as a result of the ambivalent nature of the derogation in Article 2(3) of Directive 76/207/EEC. Article 2(3) reads: 'the Directive shall be without prejudice to provisions concerning the protection of women, particularly as regards pregnancy and maternity'. This derogation acts as the connecting principle between the Equal Treatment Directive and the Pregnant Workers Directive in that the principle of subsidiarity in Article 2(3) of Directive 76/207/EEC is built on in the later Directive[24] by allowing national maternity provisions to determine the extent of positive discrimination in favour of pregnancy protection. Advocate General Darmon in *Hertz* stated that Article 2(3), in allowing Member States to adopt measures to protect women in pregnancy and maternity, was permissive rather than mandatory; thus any positive discrimination allowing for better treatment of women was at the discretion of Member States. In *Gillespie and Others* v *Northern Health and Social Services Board*,[25] *Hertz* was quoted as an example of the Court defining the perimeter of equality protection in the absence of national rules granting more favourable conditions for pregnant workers pursuant to Article 2(3). Advocate General Colomer in *Brown* confirmed the approach taken in *Hertz* and *Gillespie*. The Court of Justice has thus determined that the limits on special protection for motherhood are to be set by temporal

18 See, S. Fredman, 'A Difference with Distinction: Pregnancy and Parenthood Reassessed' (1994) 110 LQR 106, 110.

19 *Brown* v *Rentokil*, n 4 above, Opinion of Advocate General Colomer at para 64. This reasoning exposes one of the weaknesses of the 'risk' approach to pregnancy. Any protection will depend on the criterion used to assess the risk. See, further, S. Hargreaves, 'Pregnancy and Employment: a Change in European Community Law' (1995) 17 *Journal of Social Welfare and Family Law* 491, 497

20 See Fredman, n 18 above, 121.

21 Directive 92/85/EC was designed to minimise the risk of voluntary termination of employment where women have to combine child caring and working roles.

22 Directive 96/34/EC which must be implemented by December 1999.

23 Such issues have been addressed for some time in 'soft law' as recognition of the role of subsidiarity in this area.

24 See V. Cromack, 'The EC Pregnancy Directive – Principle or Pragmatism?' (1993) 15 *Journal of Social Welfare and Family Law* 261, 264.

25 Case C-342/93 [1996] ECR I-475.

© The Modern Law Review Limited 1999

barriers as fixed by national maternity laws. If this remains the case, differing social security structures across the Community clearly impede any further social advancement in favour of motherhood, unless the Court is willing to take the radical step of abandoning the alignment of protection for pregnancy-related sickness with maternity leave periods.

Advocate General Colomer hinted at a different solution to the problem in *Brown* in his summary of the domestic law provisions for the protection of pregnancy in different Member States. Although most of the jurisdictions operate a strict rule linking protection to the chronology of the illness, some Member States apply a different yardstick of capacity for work.[26] The latter rule tends to favour women's different status at work by recognising that incapacity arising from conditions relating to their reproductive role is always to be judged differently from male incapacity for work through sickness.[27]

Cost arguments

Advocate General Darmon in *Hertz* stated:

> the principle of equal treatment for male and female workers must mean a search for remedies to the problems specific to women to which pregnancy always gives rise. However, it is necessary to adjust the rules governing the labour market only in so far as the risk attaching to maternity remains one of the *'normal' risks of life*.[28]

Advocate General Colomer adopted similar reasoning in *Brown* when outlining the economic consequences of a blanket immunity for pregnancy-related sickness:

> Such a prohibition, which would affect the employer for many years after the confinement, would be liable to entail not only administrative difficulties and unfair consequences for employers but also repercussions on the employment of women.[29]

The policy reasoning which accords protection to pregnancy according to the degree of risk associated with the condition, has shaped the general approach to pregnancy protection in the European Court of Justice since *Hertz*. *Brown* has taken this a step further by crystallising the high-risk period as the pregnancy and maternity leave period.

The argument of Advocate General Colomer which equates risk with cost needs careful scrutiny. It is a classic utilitarian argument[30] that minimal risks should incur less protection where costs exceed benefits. To classify pregnancy risks by reference to time limits ie the period of pregnancy and maternity leave, has the merit of simplicity thus costs of litigation are minimised.[31] It is also cost efficient

26 The Advocate General noted that Germany, France, Denmark and Austria all apply a chronological rule whereas Spain, Finland and Sweden apply a rule based on incapacity.

27 The Swedish example might be used as a model for reform of the law in this area. Here incapacity for work does not justify dismissal if there is a probability that the worker will recover and women cannot be dismissed on account of unfitness for work resulting from pregnancy. If the 'probability of recovery' criterion were used to limit the extent of pregnancy protection, employers would not be penalised by the high costs of maintenance of unproductive employees in cases of permanent impairment.

28 *Hertz*, n 2 above, Opinion of AG Darmon, para 49 (my emphasis).

29 *Brown* v *Rentokil*, n 4 above, Opinion of AG Colomer, para 33.

30 See J. Bentham, *An Introduction to the Principles of Morals and Legislation*, ed. Burns and Hart, (London: Athlone Press, 1970) and criticisms in Sen and Williams (eds) *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982).

31 Advocate General Darmon referred in *Hertz* to the difficulties that a more flexible criterion such as incapacity might have for adjudication of claims and also to the difficulties of enforcement The problem of establishing a clear causal link would be particularly acute (see Opinion at para 47)

© The Modern Law Review Limited 1999

as more pregnancy and maternity conditions are likely to occur in this period than later on.

Advocate General Darmon in *Hertz* used this statistical argument to justify the exclusion of post-maternity complications from protection: to protect this small proportion of women might jeopardise opportunities for all women who wished to enter the labour market. This, however, appears to justify post-entry discrimination by reference to the risk of pre-entry discrimination. It also constructs the labour market by reference to a male norm as it fails to take into account large sections of the labour market which are gender segregated and reliant on female skills with typically low paid work.[32] Advocate General Colomer in *Brown* places further limits on the protection offered by Directive 76/207/EEC by indicating that the labour market must only take into account female-specific conditions during periods fixed by national law. Maternity benefits are interpreted as welfare benefits as opposed to equality rights and thus they do not confer special employment rights for women. He concludes from this that the fact that women obtain statutory maternity pay does not entitle them to any further rights:

> the simple right available to Mrs Brown who did not fulfil the necessary conditions to ... receive for a maximum period of 18 weeks a maternity benefit from the State ... does not constitute, in relation to the principle of equality, a specific employment right intended to protect the biological condition of a woman or the relationship between mother and child in the period following confinement.[33]

Since it is clearly unreasonable to expect employers to bear the cost of pregnancy and related conditions indefinitely, the ultimate question is how this cost is to be divided between employers and the State. If employers are not to be over-burdened where women's reproductive problems have reduced their productive capacity at work, a more sensitive cost indicator is needed.[34] The rule in *Brown* discounts the continuing cost of motherhood in the workplace shortly after the woman has returned to work. This solution, however, fails to take account of the cost to pregnant women who may risk unemployment if they do not return to work after maternity leave and thus remain trapped in the domestic sphere.[35]

Pregnancy and maternity rights

Recent cases have confirmed that the Court of Justice regards the rights of women on maternity leave as being quite distinct from those of pregnant women. The special protections under Directive 92/85/EC guarantee the right to maternity leave (Article 8) and an adequate allowance to maintain income during this period (Article 11).[36] The fragile compromise between equality and maternity protection

32 See J. Rubery and C. Fagan, *Occupational Segregation amongst Women and Men in the European Community* (Manchester School of Management, UMIST, October 1992): a synthesis report of the network of experts on the situation of women in the labour market.

33 *Brown* v *Rentokil*, n 4 above, Opinion of AG Colomer, para 80.

34 Directive 92/85/EC should allow more open discussion of loss distribution than Directive 76/207/EEC as health and safety relating to pregnancy cannot be fully considered without some analysis of cost factors.

35 C. Cockburn argues that women's relation to the domestic sphere is defined by means of a patriarchal sexual contract: *The Machinery of Dominance* (London: Pluto, 1985).

36 The present position in the UK is governed by the Social Security Contributions and Benefits Act 1992 and accompanying regulations under which the employer makes payments of Statutory Maternity Pay for fourteen weeks of maternity leave but recovers the cost of payments by deducting it from national insurance contributions. The rebate varies according to the size of the employer: generally employers are reimbursed at 92 per cent of payments made but 'small' employers can recover the full amount.

© The Modern Law Review Limited 1999

in the Pregnant Workers Directive, which aims to protect the condition of pregnancy and maternity as a special risk without prejudicing women's position in the labour market or derogating from the principle of equal treatment, always carried the possibility that it would further disadvantage women's employment status. The Court has repeatedly used instrumental reasons as the basis for protection of pregnancy and maternity, with the consequence that paternal ideologies have emphasised women's vulnerability rather than their rights to be treated equally as workers.[37] Despite the injunction in Article 2 of Directive 92/85/EC that the protection of the health and safety of such workers should not jeopardise equal treatment, the Court of Justice has firmly rejected the possibility of any equality comparison between women on maternity leave and other workers on the basis that such women are in a special position, but are not actually at work; thus the usual obligations deriving from the contract of employment do not arise. Furthermore, the delineation of women on maternity leave as dependant mothers as opposed to productive workers has allowed the Court to by-pass notions of equality and locate the problem in the realm of social protection where economic justifications are more likely to gain recognition. Protected status has thus resulted in a regime which has penal consequences for women in that entitlement to adequate income and immunity from dismissal during maternity leave without the need for inappropriate male comparisons has been substituted for substantive equality.[38]

The Court has shown no signs of departing from the distinction made in *Gillespie* v *Northern Health and Social Services Board* between women claiming benefits on maternity leave and men or women actually at work. The ruling out of any comparison on the basis of women's protected status in this case resulted in women's maternity benefits being calculated on the principle of adequacy of income for the protection of maternity leave rather than a worker's entitlement to full pay. Advocate General Iglesias offered an economic rationale for this: Article 2(3) of Directive 76/207/EEC envisaged subsidiarity rather than harmonisation in matters of pay, thus national expediency in matters of social security governed women's rights on maternity leave.[39]

The logic of *Gillespie* was reiterated in *Boyle* v *Equal Opportunities Commission*,[40] where the Court of Justice again rejected a comparative approach, this time between a woman on maternity leave and a man or woman on sick leave. The Court interpreted the discrimination provisions in the light of the special protections afforded by Directive 92/85/EC and deemed these protections sufficient to safeguard women's vulnerability during maternity leave in the absence of an effective comparison. The applicants in *Boyle* challenged certain provisions of a civil service maternity scheme, one of which stated that women were entitled to maternity pay in excess of Statutory Maternity Pay only if they stated that they intended to return to work after maternity leave and be liable to

Statutory sick pay used to be subject to the same principle of recoupment but the Statutory Sick Pay Acts of 1991 and 1994 have now placed the onus on the employer with the result that it is now the employer who (subject to small exceptions where his liability for sick pay exceeds liability to pay NI contributions) pays out SSP.

37 See in particular, the reasoning of Advocate General Darmon in *Hofmann* v *Barmer Ersatzkasse* Case C-184/83 [1984] ECR 3047 and commentary by C. Kilpatrick in 'European Regulation of the Post Birth Period' in T. Hervey and D O'Keeffe (eds), *Sex Equality Law in the European Union* (Chichester John Wiley, 1996).

38 See H. Fenwick, 'Special Protections for Women in European Union Law' in T Hervey and D. O'Keeffe, *ibid.*

39 *Gillespie*, n 25 above, Opinion of Advocate General Iglesias, paras 48 and 52.

40 Case C-411/96, 27 October 1998.

© The Modern Law Review Limited 1999

repay any payment if they failed to return. They contended that this clause was discriminatory in that workers on sick leave were not required to repay their contractual sick pay in the same way. The Court followed the approach taken in *Gillespie*, but achieved the same effect through a literal construction of the provisions of Directive 92/85/EC rather than relying on Article 2(3) of Directive 76/207/EEC. A distinction was made between 'pay' and 'allowance' under Article 11(2)(b) of Directive 92/85/EC, enabling the Court to differentiate between income derived from the employment relationship and income not so paid. By referring to the concept of adequacy of income under Article 11(3) of Directive 92/85/EC,[41] the Court limited the extent of maternity benefits to that provided by social security rather than the higher amount under the contractual scheme. As Article 11(2)(b) only required maintenance of an adequate allowance during maternity leave, the repayment clause was valid as the level of Statutory Maternity Pay did not fall below the income available for absence on grounds of ill health. The Court also rejected a claim that it was discriminatory to back-date maternity leave if a woman gave birth during a period of pregnancy-related sick leave. In this case, equal treatment was subsumed by the subsidiarity principle in Article 2(3) of Directive 76/207/EEC and mirrored in Article 8 of Directive 92/85/EC, allowing national legislation to determine the commencement dates of maternity leave.

The result of *Boyle* and *Gillespie* is that women on maternity leave are left without effective recourse where employers exploit the modicum of protection provided by Directive 92/85/EC. Financial detriments will continue to be incurred by women who choose to combine work and childbearing and employers will not be penalised for minimising their costs as long as the threshold of adequacy of income is not undermined.[42] These cases indicate that the limits of maternity protection are determined by the European Court's perception of national autonomy in matters of social welfare. The result of balancing competing interests on the social plane is that the cost of applying the principle of equal treatment militates against the vindication of individual rights.[43] As Advocate General Iglesias noted in *Gillespie*, to give full protection to pregnant mothers 'would threaten to upset the balance of the entire social welfare system'.[44]

By contrast with these decisions on maternity rights, recent cases concerning pregnant workers indicate that a more favourable regime applies to the pregnant woman *qua* worker. In *Thibault*,[45] the Court of Justice stated that pregnancy and maternity require substantive, not merely formal rights. Here it was the woman's capacity as a worker which the Court emphasised as the factor which triggered the operation of discrimination law, as pregnancy itself could not act as the cause for any adverse treatment in working conditions. Thus the deprival of the right to an

41 Article 11(3) provides that an allowance is deemed adequate if it guarantees income at least equivalent to that which the worker would receive if she were absent on sick leave, subject to any ceiling laid down under national legislation.

42 A recent example of the deleterious effects of *Gillespie* is *Edward* v *Derby City Council*, EAT/1353/97, 15 July 1998 (unreported) where the Employment Appeal Tribunal held that a school teacher was not discriminated against when she received half pay during a school holiday that coincided with her maternity leave

43 See C. McCrudden, 'The Cost of Removing Discrimination from Social Security Schemes' in J. Dine and B. Watt (eds), *Discrimination Law, Concepts, Limitations and Justifications* (London and New York: Longman, 1996). The argument is that a utilitarian calculus is required under the principle of proportionality in the consideration of discrimination in social security schemes, thus narrowing the scope of the right.

44 *Gillespie and Others*, n 25 above, Opinion of Advocate General Iglesias at para 48.

45 *Caisse Nationale D'Assurance Vieillesse des Travailleurs Salaries (CNAVTS)* v *Thibault* Case C-136/95 [1998] ECR I-2011.

© The Modern Law Review Limited 1999

annual performance assessment as a result of absence on maternity leave was regarded as unfavourable treatment. Notably, *Gillespie* was referred to in order to support the conclusion that the taking of maternity leave should not deprive a woman of any benefits from pay rises due to her as a worker.[46]

The bifurcation in the Court's jurisprudence on pregnancy and maternity rights has received more specific confirmation in *Hoj Pedersen*.[47] The applicants were deprived of full pay when they became unfit for work because of pregnancy-related illness before the beginning of their maternity leave. According to Danish national legislation, pregnancy-related incapacity for work entitled the worker to half pay, whereas incapacity for other illness resulted in full benefits. The Court of Justice held that it was contrary to Article 119 of the Treaty of Rome and Directive 75/117/EEC to withhold full pay from a pregnant woman who became unfit for work before the commencement of maternity leave by reason of a pathological condition connected with pregnancy.

The importance of this case lies in the explicit comparison made between pregnancy and maternity rights. Advocate General Colomer had made clear in *Boyle* that there is no basis for comparison between a woman on maternity leave and a man on sick leave because the man's contractual obligations are only suspended for so long as he undergoes treatment for a sickness, whereas maternity leave is a special 'block' period granted to protect specific pregnancy and maternity needs ie to protect the biological condition of pregnancy and the relationship between mother and child.[48] This distinction was developed in *Pedersen* where the Advocate General rejected the view that unfitness for work caused by pregnancy before maternity leave began amounted to 'pre-maternity' leave. He stated that maternity leave differs from pregnancy-related sickness absence in a number of respects: first, maternity leave has a finite duration; secondly, entitlement is not linked to illness; thirdly, a woman is released from all employment obligations including work.[49] This reasoning focuses on differences in employment rights: whereas pregnant workers must prove unfitness for work in order to be released from employment obligations, workers on maternity leave acquire rights independently of any such obligations. The dichotomy between maternity leave as social protection and pregnancy unfitness as worker incapacity allows the 'sick-male comparison' to operate only in cases of pregnancy. Gender specific rights attaching to maternity leave are diluted because they are welfare rights; pregnancy rights as equal rights emphasise the point of similarity with men and are thus conditioned by factors shaping the rights of male workers.

Pedersen also highlights the inherent instability of the chronological approach to pregnancy adopted in *Hertz* and *Brown*. The Court in *Pedersen* distinguished pathological complaints from routine pregnancy-related inconveniences. As only serious conditions result in incapacity for work, these conditions gain protection whereas routine complaints do not. This distinction, a logical result of basing protection on the concept of capacity for work, could both limit and expand pregnancy rights in the future. It will filter out minor complaints in the pregnancy period but it could also open up the possibility of allowing more serious claims in the post-maternity period. The Court used causation arguments to justify the non-protection of routine conditions by suggesting that adverse treatment resulted from

46 *ibid*, Opinion of Advocate General Colomer at para 37

47 Case C-66/96, *Handels-og Kontorfunktionaerernes Forbund i Danmark* v *Faellesforeningen for Danmarks Brugsforeninger*, 19 November 1998

48 *Boyle*, n 40 above, Opinion of AG Colomer at para 41.

49 *Pedersen*, n 47 above, Opinion of AG Colomer at para 41

© The Modern Law Review Limited 1999

a woman's choice not to work, thereby discounting the special nature of pregnancy as a female condition. The corollary of the Court's emphasis on incapacity in *Pedersen* however, is that the chronological approach to pregnancy risks in *Brown* becomes less tenable. Post-confinement complications of the *Hertz* type are also pathological complaints and should be specifically protected.

There is an implied acceptance in the judgments in *Gillespie*, *Hertz* and *Brown* of economic justifications for inferior treatment of pregnant women. In *Pedersen*, the European Court was faced with an explicit argument based on loss distribution in relation to the costs of pregnancy and firmly rejected this on the premise that discrimination cannot be justified on such grounds. The employers had argued that the discriminatory treatment was justified on the grounds that Danish legislation reflected a sharing of risks and costs of pregnancy between the pregnant worker, the employer and society as a whole. The Court stated that the aim of loss-sharing was not an objective factor within the meaning of previous case law:

> The discrimination cannot be justified by the aim of sharing the risks and economic costs connected with pregnancy ... That goal cannot be regarded as an objective factor unrelated to any discrimination based on sex ...[50]

The interests of the employer, in contrast with the interests of the State, require more specific economic justification than broad loss-sharing goals.[51]

The relationship between pregnancy discrimination and unfair dismissal

The practical consequences of *Brown* will be to enable women dismissed for a pregnancy-related illness during pregnancy and maternity leave to claim sex discrimination as well as unfair dismissal. The disparity in remedies under the Sex Discrimination Act and Employment Rights Act 1996 will prompt applicants to pursue both remedies. Recent cases such as *Jauad* v *Garrard*,[52] where the Employment Appeal Tribunal allowed a claim for interest on compensation for loss of earnings to be awarded under the more generous provisions of the Sex Discrimination Act 1975 and *Marshall (No 2)*,[53] highlight the continuing importance of pursuing parallel remedies.

Where the dismissal occurs after the end of the maternity period, according to *Brown*, there is no sexual discrimination if a male worker would have been dismissed for incapacity, thus the remedy will be for unfair dismissal rather than sex discrimination unless the illness arose during pregnancy and persisted during maternity leave. This is the effect of the rejection of *Larsson* in *Brown*. It is automatically unfair to dismiss for pregnancy or 'any other reason connected with pregnancy' under section 99 Employment Rights Act 1996 (ERA). This provision must be interpreted in the light of Directive 92/85/EC, which prohibits such dismissals up to the end of maternity leave. It can be assumed from the reasoning in *Brown* that the automatic unfairness rules will also only apply to the limited maternity leave period.

50 *ibid*, Judgment, at para 40.

51 See *Lewark* Case C-457/93 [1996] ECR I-243 where a distinction was made between economic grounds and social policy grounds of justification, the latter being more amenable to arguments based on loss distribution aims.

52 EAT/1375/96, 25 November 1997 (unreported).

53 *Marshall* v *Southampton and South West Hampshire Area Health Authority (No 2)* [1993] 4 All ER 586.

© The Modern Law Review Limited 1999

A contrary view was however recently expressed by the Employment Appeal Tribunal in *Caledonia Bureau Investment & Property* v *Caffrey*.[54] It was held in this case that the prohibition on dismissals under section 99(1)(a) ERA applies *after* maternity leave has expired, so long as the contract of employment has been expressly extended. Thus a woman who was dismissed when she suffered post-natal depression was held to be unfairly dismissed under the automatic unfairness provisions despite the fact that she had not extended her maternity leave (which would have allowed her to qualify under section 99(3) ERA). The effect of this decision would be to extend the automatic protection for post-pregnancy illnesses indefinitely. It was this spectre of indeterminate liability for post-pregnancy conditions which led the European Court of Justice to impose a finite period of protection in *Hertz* and *Brown*. The difficulty with *Caffrey* is that, even on a purposive view of section 99(1)(a), post-natal depression is more aptly characterised as maternity than pregnancy-related. However, such conditions need more adequate protection in view of the ambit of Directive 92/85/EC which is designed to protect the health and safety of workers who have 'recently given birth or are breast feeding'.[55]

Where the automatic unfairness provisions under section 99 ERA do not apply, a woman dismissed while suffering from post-natal illness will find that a male sickness comparison will be applied under the ordinary rules of unfair dismissal. The reasoning in *Brown* would imply that the purposive approach taken by the Employment Appeal Tribunal in *Caffrey* would not be followed, however continuing uncertainties in this area will necessitate extra vigilance on the part of employers who could still risk claims of sex discrimination and automatic unfair dismissal. One consequence of the continued adherence to a sick-male comparison after maternity leave has expired is the possibility of dismissals being classified as disciplinary dismissals on grounds of illness;[56] thus where there is any doubt as to the authenticity of the sickness, especially in cases of post-natal stress, employers may resort to dismissing for misconduct. This will entail the close monitoring of attendance and properly kept medical records.

A further consequence of the limited special protection provided by *Hertz* and *Brown* is that women whose pregnancy ailments persist beyond normal maternity periods, will have to negotiate the subtleties of maternity leave provision in order to claim unfair dismissal. Unless complex notice provisions are satisfied,[57] protection is nullified. In the twin decisions in *Kwik Save Stores Ltd* v *Greaves* and *Crees* v *Royal London Mutual Insurance Society Ltd*[58] the Court of Appeal applied a purposive construction of the notice provisions and held that a sick employee is not required to physically attend work on the notified day of return in order to exercise their rights. The employers had argued that the rights were lost when the employees failed to return to work, in effect using the sickness as a reason to deny the right to return and the corresponding remedy.

These cases indicate the importance of robust maternity protection if employers are not to take advantage of continuing temporary illnesses of female employees in the post-maternity period. The limitations of current protection of pregnancy and maternity illness are seen most acutely in cases where young mothers are exposed

54 [1998] IRLR 111.

55 *Brown* v *Rentokil*, n 4 above, Judgment, para 18

56 The evidence in *Brown* received adverse comment by Advocate General Colomer who noted the 'not ... very cogent reasons for ... repeated certified unfitness for work' (*ibid*, Opinion at para 19).

57 Current provisions on notification of return to work after maternity leave are contained in sections 80 and 82 of the Employment Rights Act 1996.

58 [1998] IRLR 245.

© The Modern Law Review Limited 1999

to extra surveillance at work in the period following maternity leave when they are trying to adjust to a constantly shifting divide between family and work.

Reconciliation: a new strategy for parenthood?

Recent EC policy has shifted from equality concepts to a wider definition of the work and family nexus in an attempt to integrate equal opportunities in a more flexible labour market.[59] A key theme in the *1998 Employment Guidelines* adopted by the Council of Europe[60] is the reconciliation between work and family life in Europe. The object of reconciliation is to encourage full participation of women and men in a flexible labour market by providing policies which enable parents to balance the conflicting demands of work, family and household roles. A fundamental reassessment of gender roles is taking place against a background of workplace restructuring to accommodate new demographic trends.[61] The impetus for this change of direction came from the EC Commission's Medium Term Social Action Programme 1995–1997[62] which highlighted the need to provide more flexible working hours and affordable childcare facilities to enable men and women to combine work and family life.[63] A new statutory framework for maternity and parental leave is proposed as part of a comprehensive 'family-friendly' policy in the Government White Paper, *Fairness at Work*.[64] The proposals, which aim to synchronise rights to maternity and parental leave and the corresponding rights to return to work, indicate some attempt to recast the employment relationship by allowing for longer career breaks for both males and females.

Tackling equality issues in this wider context will help to redraw the boundaries between public and private spheres by creating a more supportive environment at work. If, as a result of *Brown*, women who have not recovered from pregnancy still remain vulnerable to dismissal at work in the post-maternity leave period, this will have a harmful effect on the combination of working and caring roles which these initiatives are aiming to promote. Seen in this light, it is regrettable that the Court of Justice has provided only the minimum of protection necessary to ensure that women are not dismissed while still pregnant. However, now that *Brown* has limited special pregnancy protection to the immediate maternity period, it appears that any further rights in the post-birth period are to be accorded by means of the alternative strategy of enforcing more flexible working patterns. Advocate General Colomer was perhaps merely stating the obvious when paraphrasing Papinian: 'despite the passage of centuries "there are many points in our [Community] law in which the condition of females is inferior to that of males" '.[65]

59 See B. Bercusson, *European Labour Law* (London. Butterworths 1996) ch 15.

60 1998 Employment Guidelines adopted by Council of Europe, 15 December 1997.

61 See the Report on *Reconciliation between Work and Family Life in Europe* by the European Commission's Gender and Employment and Gender and Law groups of experts, 1998, *Equal Opportunities for Women and Men in the European Union*, Annual Report, 1997, European Commission, Luxembourg.

62 COM(95) OI34-C4-0160/95, Resolution on the Medium-Term Social Action Programme 1995–97, Brussels, 12 April 1995.

63 The concept of reconciliation is developed in more depth in the Consultative Document, *Reconciliation of Professional and Family Life*, European Commission, Luxembourg, February, 1995.

64 Cm 3968, May 1998

65 *Brown* v *Rentokil*, n 4 above, Opinion of AG Colomer, para 67.

© The Modern Law Review Limited 1999

Doctrinal Rationality after *Woollin*

William Wilson*

There will be some who breathe a sigh of relief following the House of Lords' decision in *Woollin*.[1] At last, it seems, the meaning of intention in the criminal law is cogent and stable and the mens rea for murder is as close as the common law is likely to get in providing a bright line between the culpability associated with this crime in contrast with manslaughter. The state of the law seems now largely to coincide with the meaning optimistically attributed to it by the majority of academic commentators and successive reform bodies.[2] The fudge, initiated in *Hancock*[3] and which reached its apogee in *Nedrick*[4] as to whether intention is a state of mind which can be inferred from foresight of probability or, if not probability, then virtual certainty, or whether foresight of virtual certainty is a form of intention has been eradicated. The House of Lords in *Woollin* has firmly nailed its colours to the mast of the latter. Everything has fallen neatly into place. Or has it? I hope to show here that it has not – that the present meaning of intention is still generally unstable and that the decision renders the broad doctrinal terrain of homicide less intelligible.

The facts

W, having lost his temper with his baby son, threw him with much force across the room causing the infant to hit his head on a hard object. The latter later died as a consequence of his injuries. W was charged with murder. Under interrogation W admitted he realised there was a risk of serious injury but denied any desire to kill or hurt the child. The prosecution case, relying on *Nedrick*, was that the necessary intention to kill or cause serious injury was present if, when acting, W must have known it was virtually certain that such injury would be caused. His defence, *inter alia*, was lack of intent to cause serious injury.

The trial judge directed the jury that they must be satisfied W intended the death of or serious injury to the child. At first, he directed them, in accordance with the model direction in *Nedrick*, that they could infer intention from their belief, if they had one, that the accused foresaw either of these consequence as a virtual certainty. However he went on to say:

> If he had not given any thought to the consequences of what he was doing then the Crown would have failed to prove the necessary intention to cause really serious harm and you should acquit. On the other hand if you reject that interpretation and were quite satisfied that

* Law Department, Brunel University.

1 [1998] 4 All ER 103.

2 See for example A.J. Ashworth 'Criminal Liability in a Medical Context: The Treatment of Good Intentions' in A.P. Simester and A.T.H. Smith (eds) *Harm and Culpability* (Oxford: Clarendon, 1996) 173; JC Smith 'A Note on Intention' [1990] Crim LR 85; J.C. Smith and Brian Hogan *Criminal Law* (London: Butterworths, 1996) 58–60; Law Commission *A Criminal Code for England and Wales*, No 177 cl 18, (London: HMSO, 1989).

3 *Hancock and Shankland* [1986] 1 All ER 641.

4 [1986] 3 All ER 1.

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

> he was aware of what he was doing and must have realised and appreciated when he threw the child that there was substantial risk that he would cause serious injury to it then it would be open to you to find that he intended to cause serious injury to the child.

The jury convicted on this basis.

Woollin in the Court of Appeal

It was argued on appeal that the judge had misdirected the jury. The approved direction in cases where the jury needed guidance as to what intention meant was that appearing in *Nedrick*. This is as follows:

> Where the charge is murder and in the rare cases where the simple direction ... is not enough, the jury should be directed that they are not entitled to infer the necessary intention unless they feel sure that death or serious bodily harm was a virtual certainty (barring some unforeseen intervention) as a result of the defendant's actions and that the defendant appreciated that such was the case.[5]

The basic defence objection was that the jury might have thought that '[realising that] there was a substantial risk that he would cause serious injury' *constituted* the intention demanded by the offence definition. Clearly this is not the case and has not been so since *Moloney* effectively overruled *Hyam*.[6] In this latter case, on a charge of murder, a majority of their Lordships had decided that a person who acted with foresight of a consequence (here death or serious injury) to a specified (though uncertain) degree of probability intended that consequence. This ruling both blurred the distinction between murder and manslaughter and between intention and recklessness. *Moloney*, in an attempt to recover some doctrinal clarity, stated that what a person foresees and what he intends are different states of mind. The former could only, therefore, operate as evidence of the latter.

The Court of Appeal, in *Woollin*, said that the trial judge's direction was acceptable since it had been made clear to the jury that they were looking for intention and that foresight of a substantial risk was only evidence from which they might infer such intention. The view was taken that the jury would not have misunderstood from the direction as a whole that the judge was asking them to find intention. This finding would have been more plausible perhaps if the trial judge had defined intention. But he did not. As a result we have no means of knowing whether the jury convicted because they thought that foresight of substantial risk was enough in itself, or because, on the basis of the evidence put to them, they were convinced that it was the accused's purpose to cause the baby serious injury, or because they favoured some other test of intention – say an attitude of utter indifference. It is difficult to see, then, how the Court of Appeal could have reached the conclusion that the jury would not have mistaken what they were supposed to be looking for. Frustratingly, as with the trial judge and following the example set in *Hancock* and *Nedrick*, the Court of Appeal did not specify what intention meant and how foresight of the substantial risk related to it.

Nedrick had offered the following (equivocal) insight. Foresight of a consequence is certainly evidence that it was intended if it is foresight of virtual/moral certainty. If F throws his son off the top floor of the Empire State Building he can deny the intention to kill until he is blue in the face, but nobody need take

5 Cited at [1998] 4 All ER 103, 105.
6 *Hyam* v *DPP* [1974] 2 All ER 41, *Moloney* [1985] 1 All ER 1025.

© The Modern Law Review Limited 1999

the claim seriously unless, perhaps, the outcome (death or serious injury) was an outcome he was taking this action to avoid.[7]

Foresight of virtual certainty may perhaps, though, be more than *evidence* of intention. It may be a separate *form* of intention.[8] Lord Lane CJ was keen to hedge his bets in *Nedrick*. If so, it will make no legal difference why F threw his son from the building, although clearly it may make a moral difference. One way or another the jury are entitled to find the intention (proven). They need not trouble to consider why he did it. So also, a wicked surgeon who removes another's liver for transplant into his dying wife may not desire the former's death either for its own sake or as a means to achieve his end but he would intend it if he knows it is practically certain, barring some unforeseeable intervention by another surgeon with a spare liver.

Nedrick has been widely criticised for not making clear whether or how foresight of probability/high probability falling short of virtual/moral certainty may be *probative* of intention. It is only clear, in line with *Moloney* and *Hancock*, that such foresight does not *constitute* intention. Logically, however, in many cases it will not even be probative. If, for example, the wicked surgeon removes only a kidney he may, if he wantonly fails to adhere to orthodox surgical procedures, foresee the victim's death as probable or even highly probable, but this is not evidence that he means her to die.[9] His true motive shows this beyond any question. It is evidence – the strongest imaginable – that he does not care whether she lives or dies but not evidence of intent unless intent carries its pre-*Moloney* meaning. This is the bright line which Lord Lane CJ tried to draw between intention and recklessness. It can only be such evidence where the facts show that there was no room for the actor to doubt that the consequence would not occur and so there is no room also for the jury to doubt that he did not decide to produce the consequence. As soon, in other words, as the consequence becomes less than morally certain doubt must creep into the jury's collective mind as to whether it was meant.[10]

How, then, did the Court of Appeal justify their decision in *Woollin*? Unpersuasively, it must be admitted. At least for this commentator, attempting to divine the merit of the court's reasoning necessitated the unusual step of long deliberation with eyes shut in a darkened room. The best way to cast light on its overall cogency is to apply it to some straightforward examples. Roch LJ said that the *Nedrick* direction was only necessary in cases where the sole evidence of D's intention is evidence of D's actions and their consequences to the victim.

Case 1

> A is found dead at the foot of a block of flats. B was seen pushing him from the window of the fifth floor. B offers no evidence as to why he did it or his state of mind at the time of doing so.

7 As where the building is on fire and A throws the deceased from the building in a vain attempt to save life. See Law Commission, *Legislating the Criminal Code: Offences Against the Person and General Principles* No 218 (London: HMSO 1993). For further discussion see below

8 By analogy, though we appreciate that courgettes belong to the family of fruits they are nevertheless designated, for gastronomic purposes, as a *form* of vegetable.

9 Neither, of course, will it constitute the form of intention conveyed by foresight of moral certainty. It might conceivably corroborate an *independent* malicious motive for killing the 'donor'.

10 For an extended discussion of this point see W. Wilson *Criminal Law: Doctrine and Theory* (London: Longman, 1998) 126–8.

© The Modern Law Review Limited 1999

This is a case where the judge, in accordance with *Nedrick*, would tell the jury that in the absence of a motive, or other evidence of what B had in mind it was open to them to infer that B intended to kill A only if they were sure that such a consequence was virtually certain and that B knew this.

Roch LJ said that if there is other evidence then the jury may be invited to consider all the evidence and from that evidence draw an inference of intent, in accordance with section 8 Criminal Justice Act 1967.[11] This would be so even though the consequence was not virtually certain and was not foreseen as such.

Case 2

> As in Case 1 except that A was dropped from the first floor and B had been heard earlier to threaten to kill A. B had admitted that he knew there was a substantial risk that A might suffer serious injury but denied intending to kill or cause serious injury.

This is a case where the judge should tell the jury that they must determine intention by reference to all the evidence, including presumably (a) the earlier threat, (b) the admitted foresight, (c) the objective dangerousness of B's conduct as indicated by the height from which A was dropped. This would seem to be a permissible departure from the spirit of *Nedrick* since there is evidence that B wanted to kill A. Given that there was a substantial risk that death or serious injury would occur and he was alert to this fact his conduct corroborates that evidence. It goes without saying that his admission that he foresaw the consequence as probable would not be probative of the fact that he foresaw it as morally certain unless, objectively, it was.

Case 3

> As in Case 2 except that there is no evidence of B's state of mind other than his admission that he foresaw the substantial risk.

Case 4

> As in Case 3 except that B gives, as his explanation, the fact that he wanted to teach A a lesson by frightening him.

In Cases 3 and 4, since the evidence relied upon by the prosecution is not limited to B's actions and their consequences to the victim, the judge may direct the jury to determine B's intent by reference to all the evidence, which will include the fact that he foresaw the consequence as a substantial risk in Case 3 and also wanted to teach A a lesson in Case 4.

As explained above the major problem posed by the direction of the trial judge in *Woollin* is that the jury were not told what they were supposed to be looking for when determining B's intent. They were not told what this intention is whose existence may be inferred from B's foresight.[12] Is it aim or purpose? The jury were told that they were not looking for this type of intention in *Woollin*. In theory, however, what a person foresees is not necessarily even probative of what he

11 Which provides that what an actor intends or foresees must be determined by the jury on a consideration of all the evidence.

12 The strong possibility is that they will infer nothing at all, but treat such foresight as constitutive of the fault element of the offence. See Sir John Smith's commentary on *Woollin* at [1997] Crim LR 519, 520–1.

© The Modern Law Review Limited 1999

means to achieve. Direct intention and foresight are different states of mind, in the same way that love is different from acquisitiveness. Proving that a person foresees a consequence as probable/highly probable is no more conclusive of an intention to produce that consequence than counting an art dealer's acquisitions can establish his love of art. However much the dealer acquires the possibility is that he hates art, but loves accumulating wealth. Similarly, direct intention cannot conclusively be inferred from mere foresight of probability. It may help us make a pretty good guess but guesswork is not the stuff of criminal convictions. Suggesting otherwise fatally conflates intention and recklessness. This can be seen at its most obvious in Case 4. Why should the jury prefer to believe that B intended to kill rather than to frighten given that the evidence equally supports both possibilities? The standard of proof, it must be remembered, is beyond reasonable doubt.[13]

If we include foresight of moral certainty as a separate *form* of intention it might be argued, however, that where an undesired consequence is, objectively, morally certain then an admission that the consequence was foreseen to a lesser degree of certainty is probative of intention as thus defined. In Case 4, assuming the victim was pushed from the *tenth* floor, the jury may reasonably conclude, that (a), since death under these circumstances was morally certain, and (b), since there is clear evidence that B was acting as a fully sentient moral agent – he had made a risk evaluation – then (c), B, who was not stupid, also must have known what everybody else would have known. Death was not merely a probability; it was a certainty. This is possibly what Roch LJ had in mind in seeking to draw a distinction between cases where probability of the occurrence was the only evidence the jury have to go on in deciding the accused's (direct) intent and cases where the accused appears to be denying the obvious, namely that the risk which he says he only foresaw as substantial, was a risk which the rest of the world would foresee as overwhelming.

In *Woollin* D had also admitted that he realised serious injury might thereby be caused and thus left it open to the jury to conclude not merely that he foresaw the consequence as probable but that he foresaw it as certain. As explained above, this might be a reasonable inference where the consequence *was* to all intents and purposes inevitable and must have been recognised as such. But here we do not have a person who pushes his baby out of a window but one who throws her across the room in a momentary loss of control. How could the jury conclude, beyond reasonable doubt, either the objective certainty of the consequence or, indeed, that D must have appreciated it? If they were in doubt they must give the benefit of that doubt to the accused. It should be noted keenly that the prosecution had seen fit not to contend at trial that the appellant had desired to kill his son or cause him serious injury. Given this signal fact what were the jury thinking of in finding intention proven? It seems only too obvious that they felt able, given the looseness of the direction, to reach a conclusion that D deserved to be convicted of murder without stretching themselves to find either direct intention or foresight of moral certainty.[14]

The certified questions

The Court of Appeal certified the following questions for their Lordships' consideration:

13 cf A.P Simester and W. Chan 'Intention Thus Far' [1997] Crim LR 704.
14 See A. Norrie 'Oblique Intention and Legal Politics' [1989] Crim LR 793 800–7.

© The Modern Law Review Limited 1999

In murder, where there is no direct evidence that the purpose of the defendant was to kill or inflict serious harm on the victim, is it necessary to direct a jury that they may only infer an intent to do serious injury if they are satisfied (1) that serious bodily harm was a virtually certain consequence of the defendant's voluntary act and (2) the defendant appreciated that fact? If the answer to the first question is 'Yes' is such a direction necessary in all cases or is it only necessary in cases where the sole evidence of the defendant's intention is to be found in his actions and their consequence to the victim?

Tidying up: the House of Lords' response

The House of Lords, in declining to answer the certified questions, showed themselves alert to most of these problems. They disapproved of the distinction drawn by the Court of Appeal between those cases where the *Nedrick* direction was appropriate and those where it was not. Lord Steyn said that this would make the law both complicated and unprincipled. In so doing their Lordships rejected the Crown's argument that the *Nedrick* direction was bad because, in removing foresight of mere probability from the scope of probative evidence, it was inconsistent with section 8 of the Criminal Justice Act 1967.[15] Lord Steyn's reasoning was that *Nedrick* lays down a rule of law – intention can take the form of acting in the knowledge that a consequence is certain – not a rule of evidence – (direct) intention cannot be established by adducing evidence that the actor foresaw the result to a lesser degree of probability than virtual certainty.[16] As explained above, in most cases foresight of probability will not be probative of direct intention and this is why a judge should not direct them to this effect. If he/she does so it will certainly confuse them. It may plausibly be probative of foresight of virtual certainty where the consequence is, in fact, objectively certain if things go according to plan. In such circumstances the jury may reasonably conclude 'Well if you foresaw it as likely you must have also foreseen it as certain because it was certain'.[17]

They also, less convincingly, rejected the argument that it was inconsistent with *Hancock*. Lord Scarman had said in *Hancock* that the more probable the consequence the more likely it was foreseen and, if foreseen, the greater the probability that it was intended.[18] This can only mean, it is suggested, that what A wants to happen (direct intention) may be inferred from what he does and as the probability of that consequence occurring as a result of his conduct increases so also diminishes the jury's room for doubting whether A does in fact want it to happen. A juror who has no other evidence of what A wanted to do other than that he dropped B out of a high window will no doubt wish and, by dint of this remark, be emboldened to infer that A wanted to achieve that which his conduct made highly probable.

Lord Steyn, giving the leading opinion, by no means did justice to Lord Scarman's remark. He might reasonably have conceded that such foresight is probative of *direct intention*[19] where, but only where, D has no plausible

15 In cases where there was other evidence of A's state of mind and see footnote 11.

16 This is not what Lord Steyn said, but it is the interpretation which makes best sense of what he said.

17 cf Smith and Hogan n 2 above, at 58 who insist that this form of intention requires only that D foresees the consequence as certain; it should not be necessary also that the consequence is, objectively, certain. The example they give to support their conclusion (a shooting) is one where the special direction would not be appropriate.

18 at 651.

19 Foresight, not foreseeability, is the point here as Woollin admitted he had foresight. Typically however a defendant will make no such admission, which is where Lord Scarman's guidelines would kick in. See previous paragraph.

© The Modern Law Review Limited 1999

explanation for his behaviour other than that he desired the proscribed consequence. In such cases it seems unobjectionable, if perhaps unwise, for judges to point out something which common sense will compel them to do anyway, namely to decide what a person wanted to achieve by reference to what he thought would probably happen.[20]

Significantly, in such cases Roch LJ had said that the *Nedrick* direction is applicable. This view is particularly hard to comprehend. The special direction should only be necessary where the jury might lose sight of what mental attitude they were looking for. If the defendant offers no competing view of his conduct why should one be necessary? Beyond this Lord Steyn might also have conceded that foresight of less than certainty is probative of *oblique intention* (foresight of certainty) but only if the consequence *is* objectively certain, barring some unforeseen intervention.

The House concluded that the judge should not have departed from the *Nedrick* model direction which was 'a tried and tested formula'[21] designed to draw a distinct line between intention and recklessness and between murder and manslaughter. Lord Steyn located the main source of the confusion in *Nedrick* as Lord Lane CJ's attempt to do justice to Lord Scarman's 'taking inferences from probability' statement in *Hancock*. This perpetuated the confusion as to whether knowledge of virtual certainty was a form of intention or was (possibly) irresistible evidence of (direct) intention[22] and also left it unclear whether, if the latter, foresight of a lesser degree of probability was still probative. This part of the *Nedrick* guidelines should not now be put to the jury as it is confusing. Although it may sometimes be necessary for juries to be given guidance on how foresight of probability may relate to proof of intention, this should be exceptional. When it is given it should be in the form of the clear statement of Lord Lane CJ in the model direction with one qualification, namely substituting for 'infer' the word 'find'. It should conclude with the rider that 'the decision is one for the jury to be reached upon a consideration of all the evidence.'

The special direction, with modifications emphasised, would, then, read:

> Where the charge is murder and in the rare cases where the simple direction is not enough, the jury should be told that they are not entitled to *find* the necessary intention, unless they feel sure that death or serious bodily harm was a virtual certainty (barring some unforeseen intervention) as a result of the defendant's actions and that the defendant appreciated that such was the case. *The decision is one for the jury to be reached upon a consideration of all the evidence.*

Interpreting *Woollin*

There are two ways of reading this adoption and modification of the *Nedrick* direction. The first is that the House of Lords thought they were implementing the standard test of intention which holds that consequences of action are intended if they are either (a) desired or (b) foreseen as a virtual certainty. The second is that the House of Lords consciously declined to implement the standard test, presumably in order to perpetuate existing (desirable) doctrinal ambiguity.

Reading *Woollin* as a whole the most cogent interpretation is the former and is

20 And so decided the Court of Appeal in *Walker and Hayles* (1990) 90 Cr App R 226.

21 [1998] 4 All ER 103, 113, per Lord Steyn.

22 Or possibly some other notion of intention. We have no means of knowing as intention was not defined.

© The Modern Law Review Limited 1999

the one taken here. The clearest expression of the standard definition is in the form proposed by the Law Commission and is intended largely as a restatement of the existing law. It states that a person acts 'intentionally' with respect to a result when (i) it is his purpose to cause it; or (ii) although it is not his purpose to cause that result, he knows that it would occur in the ordinary course of events if he were to succeed in his purpose of causing some other result.

The definition renders it clear that a person intends, for example, to kill both when he wants to kill and also when he recognises that death is a morally certain side effect of what he does want to achieve if things go according to plan.[23] Foresight of virtual certainty is a form of intention not simply a means of proving intention. It designates the wicked surgeon's conduct as intentional with respect to the death of his patient if he removes her liver but not if he removes her kidney.

This interpretation, though the most cogent, may yet be surprising since the direction is couched in the conditional negative; it does not say '*you must* find intention *if* the defendant foresaw the consequence as certain;' rather it says '*do not* find intention *unless* the defendant (desired the consequence or) foresaw the consequence as certain'. It may thus be understood to imply that while foresight of certainty is a precondition of this form of intention it is not reducible to it. The major ground for rejecting this implication is the use Lord Steyn makes of the complaint of Ashworth, J.C. Smith and Glanville Williams that the use of the word 'infer' detracts from the clarity of the model direction. This leads him to conclude that 'find' is the *mot juste*. Each of these commentators gave as their reason for rejecting 'infer' on grounds of 'clarity' the fact that the only (intending) state of mind which could meaningfully be inferred from foresight was direct intention. As direct intention was something other than the state of mind under consideration the latter must be an alternative *form* of intention, rather than a means of *proving* intention.[24] This reasoning has been accepted subsequently by the articulator of the *Nedrick* direction, namely Lord Lane CJ, who has conceded in the course of debate that what he would have liked to say is that a person brings about a consequence intentionally not only when he acts in order to bring it about but also when he foresees it (as virtually certain). Common sense would also suggest that in preferring 'find' to 'infer' Lord Steyn was conceding the general conceptual point underpinning this objection rather than simply articulating a more incisive form of words. At other parts of his judgment this impression is confirmed. Earlier he said '[t]he effect of the critical direction [ie the one he later adopts] is that a result foreseen as virtually certain is an intended result'.[25] At another he states '*Nedrick* does not prevent a jury from considering all the evidence: it merely stated what state of mind (in the absence of a purpose to kill or to cause serious bodily harm) is sufficient for murder'.[26] The obvious conclusion to be drawn is that Lord Steyn presumed that the *Nedrick* model direction gave as the (second) state of mind

23 This definition satisfies the objection lodged above at footnote 21.

24 By analogy, an untutored view of the cosmos holds that there are two basic forms taken by celestial bodies – those which are visible and those which are not. We find the former by means of a telescope which registers light waves. We find the latter by means of a telescope which registers radio or other waves. In both cases we use the telescope to find the body. We do not say, in the case of invisible bodies, that the radio signal allows us only to infer the existence of the body. We would use the language of inference only when it was neither visible nor otherwise directly detectable as when the existence of 'dark matter' (which sends out no calling card) is inferred by reason of the failure of the universe to collapse in on itself.

25 [1998] 4 All ER 103, 110.

26 *Ibid* at 111.

© The Modern Law Review Limited 1999

sufficient for murder the appreciation that death or serious injury was a virtual certainty (barring some unforeseen intervention) and his intention was to underline this by replacing 'infer' with 'find'.

The best argument for the second reading is that the direction means just what it says. Juries are entitled to find intention where there is foresight of virtual certainty, which is not the same as being bound to. Before discussing why the House of Lords might have desired to leave intention 'up in the air' in this way is there any evidence that they might have had this explicitly in mind? There are two possible signals. The first is that Lord Steyn rejected the opportunity of adding the following sentence to the model direction which Lord Lane had fashioned in *Nedrick* as an aid to the jury:

> Where a man realises that it is for all practical purposes inevitable that his actions will result in death or serious harm, the inference may be irresistible that he intended that result however little he may have desired or wished it to happen.

The second is that he did add, with Lord Hope's approval, the subsequent sentence namely that 'the decision is one for the jury to be reached upon a consideration of the evidence'. It is just plausible that the first sentence was rejected because it provided an unwelcome restriction on the scope available to the jury to resist finding that the consequence was intended. This hypothesis is in no way weakened by their Lordships' approval of the second sentence which also did not form part of the original model direction. Indeed it is strengthened by it. Read as a whole the direction may be seen as incorporating an explicit 'get out clause' for juries who believe the accused to have foreseen the prohibited consequence as a virtual certainty but nevertheless still feel able to conclude – by dint of a 'contextualised' usage of the word intention – that the consequence was not intended. The last sentence might be understood as the judicial wink and the nudge which, in appropriate circumstances, might embolden the jury to deploy common sense before linguistic rectitude.

Doctrinal ramifications

Whichever reading we prefer it is apparent that their Lordships, following the example set in *Moloney, Hancock and Shankland*, and *Nedrick* eschewed once more the opportunity to present a clear and cogent definition of intention. As a consequence whatever the House of Lords had in mind in approving the tidied-up direction the lack of such an unequivocal definition threatens to dog rational doctrinal development. The position now reached is that we are reminded forcefully what intention is not, namely the state of mind of someone who foresees a consequence to a lesser degree of probability than that of virtual certainty. We know, moreover, that a judge must not direct the jury that they may infer intention from foresight of probability for fear of misleading them. Intention is thus distinguished conceptually from recklessness in a way which previously it was not. This, in itself, is very much to be welcomed. The *Nedrick* guidelines were easily deployed, consciously or unconsciously, to create the impression that extreme cases of risk-taking constituted, in the context of murder at least, its fault element. If, say, intention polices the boundary between murder and manslaughter it is of key concern to ensure that this boundary is at least conceptually distinct. This is achieved where intention's meaning is limited to consequences desired or foreseen as morally certain and where there is no room for a judge to present the jury with a

 © The Modern Law Review Limited 1999

form of words which allows them to apply common sense or broader penal goals before doctrine.[27]

Beyond this negative definition, we can also be quite confident, on the basis of existing authority,[28] that an actor intends (to act or) bring about a consequence if he acts in order to bring it about either for its own sake or to achieve some other purpose. In these latter cases, it must be understood moreover that the jury does not need to be convinced that the accused knew for certain that he would succeed in his purpose. It is enough that they are sure he wanted to and acted in this way because of it. That what he did was consistent with this desire is the only possible probative value of adducing degrees of probability.

The House of Lords' limited tidying up operation stores up two obvious problems of application. The first involves two special cases of good motives ((a) and (b) below) and is directly attributable to the failure to implement explicitly the standard definition. The second, of particular significance in murder, involves cases of wicked recklessness ((c) below) and is directly attributable to the clear statement that juries should not be directed on the probative value of degrees of foresight less than that of virtual certainty. The problem cases are as follows:

(a) The defendant's motive is morally inconsistent with the criminal intention required by the crime definition. An example is *Adams*.[29] Does a doctor's quite lawful intention (pain relief for a dying cancer victim) transmute into an unlawful intention (to kill the patient) by crossing a threshold of probability (eg from highly probable to certain)? Another example is *Gillick*.[30] Here the doctor prescribed contraceptives to an under-age patient for the purpose of preventing a disastrous pregnancy. If he knows this will facilitate the commission of (unlawful) sexual intercourse with the patient does his lawful intention transmute into an intention to aid and abet the offence?

(b) The defendant's motive is conceptually inconsistent with the criminal intention required by the crime definition. An example is given by the Law Commission: F throws S off the top of a blazing block of flats in a vain attempt to save S's life. If a person acts in order to avoid the consequence does he nevertheless intend that consequence if he foresees its occurrence as inevitable.

(c) The evidence shows that the defendant knew the consequence, say death, was likely to occur by reason of his action and might, therefore, have desired it. But he has a plausible alternative reason for his action which, if true, would negate such desire but not the contempt for the value of life which the murder conviction condemns.[31] An example is *Hyam*. The defendant anxious to drive her rival out of the neighbourhood and her lover's life had put a lighted newspaper through her letter box. Two children of the household died in the resulting blaze. Mrs Hyam's conduct was consistent both with intending to kill the occupants and also with intending to frighten the rival into leaving her husband.

27 See A. Norrie 'Oblique Intention and Legal Politics' [1989] Crim LR 793 at 800–7; R.A. Duff 'The Politics of Intention: a Response to Norrie' [1990] Crim LR 637.

28 See *Mohan* [1976] QB 1.

29 [1957] Crim LR 365.

30 *Gillick* v *West Norfolk and Wisbech AHA* [1986] AC 112.

31 This mental attitude is described in other jurisdictions as wicked recklessness or depraved heart.

© The Modern Law Review Limited 1999

Good motives

One notable effect of the *Nedrick/Hancock* ambiguity is that it lent judges explicit power, in cases involving a justifying motive such as *Adams* and *Gillick*, to let the jury reach their own common-sensically humane decision. Effectively, then, case law was riven by two complementary, if inconsistent, methods of constructing the fault element designated by the word intention. In cases involving dangerous activity disclosing no social value the fault element was routinely expressed to include acting for the sake of a consequence and acting in the knowledge that a consequence was virtually certain. In cases involving a justifying motive trial judges had discretion to stress that foresight even of virtual certainty was only evidence of intention and that the jury should make their determination on a consideration of all the evidence. They could go further and give a direction which excludes the standard test of intention.[32] *Woollin*, with its preference for the word 'find' rather than 'infer' appears to have been intended to remove the possibility of doing this. Judges, we are told, must give the unadulterated (modified) model direction in any case where the basic direction is eschewed.

On further consideration the present position is not quite so straightforward, however. *Woollin* contains no explicit suggestion that either *Adams* or *Gillick* has been disapproved. Moreover, Lord Steyn made a studied effort to ensure that trial judges understand that the *Nedrick* direction is not automatically to be given, perhaps even in cases lying outside intention's focal cases:

> It *may* [my emphasis] be appropriate to give a direction in accordance with *R* v *Nedrick* in any case in which the defendant may not have desired the result of his act. But I accept that the trial judge is best placed to decide what direction is required by the circumstances of the case.[33]

The permissive rather than imperative emphasis in this statement may well mean that in cases where the accused's motive contradicts definitional intention or is morally inconsistent with it the judge will feel encouraged tactfully not to put such directions to the jury and trust in their common sense. The bolder alternative is to continue with the *Adams/Gillick* approach. In medical cases in particular, given that we know doctors in such cases are acting lawfully it is unlikely that a trial judge, with his back to the wall, will scruple to determine why, as a matter of doctrinal necessity, this is so.[34] And in cases of conceptually inconsistent intentions it would clearly not make doctrinal, semantic, or moral sense to find a criminal intention. In the example above, if F's intention is to save S's life, he cannot *as a matter of definition* intend to kill S, whatever intention might mean.[35] The two intentions are incapable of inhabiting the same mind at the same time. Everyday usage understands this very well.

Even if the judge chooses to give the modified *Nedrick* direction, moreover, it should be remembered that this does not foreclose the possibility of a linguistically sophisticated jury saying to themselves 'OK, we are entitled to find intention here but we prefer not to. People intend what they set out to achieve. They do not intend the undesired side effects of what they are in business to achieve, particularly

32 See, for example, Lord Scarman in *Gillick*: 'The bona fide exercise by a doctor of his clinical judgment must be a complete negation of the guilty mind which is an essential ingredient of the criminal offence', at 190. Also *Steane* [1947] KB 997.

33 [1998] 4 All ER 103, 112.

34 Of course problems will still arise in cases outside the medical arena.

35 The Law Commission's definition of intention expressly covers the case of contradictory intentions and would render F's conduct non-intentional as to the death of S.

© The Modern Law Review Limited 1999

where these are inconsistent with the point of the defendant's action'. Neither, more importantly, does it foreclose the possibility of a similar jury embarrassing the trial judge by asking for a further direction on this very point. How should a judge respond to such an inquiry? The real weakness of *Woollin* lies in its resounding silence on this issue. There are two obvious options available to the judge.

The first is to bat the question of meaning back to the jury by telling them that intention is a word with an everyday usage and it is that usage which they must deploy. It goes without saying that in these grey areas where everyday usage lacks clarity and focus such a response would be a dereliction of duty. It would similarly be a dereliction of duty to avoid giving a special direction in the first place which has always been an option for a trial judge.

The second is for the judge to grasp the juridical nettle and tell the jury that if they find that the accused foresaw the consequence as certain then they must also find that he intended it. But this (most obvious) option carries a nasty sting in the tail. The standard definition converts a person's lawful intention into an unlawful intention not by reference to any mischievous disposition or choice but simply by reference to the state of his knowledge. Knowledge as a fault element, unlike recklessness, does not filter out liability for actions prompted by good motive.[36] Risk-taking, on the other hand, is only culpable and so punishable if unjustified. The House of Lords gives trial judges no guidance as to how to plug this culpability gap and, indeed, by *Woollin*, might be thought to have exacerbated the problem. For example, A is driving quite lawfully on a narrow mountain cliff road and encounters, on turning a corner, a bunch of hikers stretched out across the road.[37] If, instead of plunging in heroic self-sacrifice off the cliff, he ploughs through the middle of them he does not act recklessly for, say, the crime of reckless driving if this was his only way of staying alive.[38] His risk-taking is justified. The criminal law imposes no duty of self-sacrifice in the interests of others. If, on the other hand, a hiker died and A is charged with murder the mental element is satisfied under the standard definition. A's liability would, then, depend upon his being able to raise a justificatory defence. As the law now stands, however, there is no such defence[39] and a trial judge is in no position to invent one.

This problem is seen at its most intractable in the field of accessories.[40] A gunsmith who supplies a gun, under licence, to a customer is acting lawfully. If, however, he knows that the customer intends to use it to commit a criminal offence he is deemed, under the standard meaning, to intend to assist the commission of the offence. A lawful intention – to ply one's trade – becomes unlawful not by reference to the gunsmith's reasons for acting but by reference to his knowledge – knowledge it should be noted that he has no duty to do anything with. He need not inform the police of the customer's plans. He may sell him a holster, or a ten gallon hat. His autonomy is respected this far. But no further. He may not sell him the gun. In standard cases such as this we may not scruple to sympathise too much. He is acting anti-socially after all. But doctrine must fit the non-standard case as well and when it does, as where, say, M gives her head-strong under-age daughter,

36 See generally A.P. Simester 'Why Distinguish Intention from Foresight?' in Simester and Smith (eds) n 2 above

37 For discussion of this hypothetical see W Wilson 'A Plea for Rationality in the Crime of Murder' [1990] 10 LS 307; G. Williams 'Rationality in Murder: A Reply' (1991) 11 LS 204.

38 The modern version of this offence is dangerous driving. The point at issue is not affected.

39 Duress of circumstances which most nearly covers the case is not available to murder.

40 See most recently *JF Alford Transport* [1997] 2 Cr App R 326.

© The Modern Law Review Limited 1999

contraceptives in a desperate attempt to prevent an unwanted pregnancy, doctrine is placed under strain.[41] Something has to give – doctrinal integrity or criminal justice.[42] Implementing the standard definition leaves the absence of any generally applicable justificatory defence, such as might exculpate M – a serious lacuna. Judges could get by without it so long as they had, in the definition of intention, a flexible friend which, in appropriate circumstances could allow good intentions to take doctrinal precedence over knowledge.

The broad message conveyed here is that *Woollin* leaves the problem posed by defendants with justificatory motives unsolved, whether we interpret it as instantiating the standard definition or simply as removing an undesirable source of uncertainty. In the former case unless trial judges avoid giving the special direction in morally troublesome cases[43] or confine the standard meaning to *Woollin's* own fact configuration,[44] we are in urgent need of doctrinal growth in the field of justificatory defences stretching far beyond its present medical confines.[45] In the latter case the uncertainty remains in cases of good intentions even if it seems to have been removed elsewhere. *Woollin* is by no means the end of the story.

Risk-taking

In *Woollin* the Crown did not argue that, as a matter of policy, foresight of virtual certainty is too narrow a test in murder. This is perhaps surprising given the social opprobrium attached to killings in the course of creating terror but where there is insufficient evidence of direct intention. By increasing doctrinal certainty we do not, then, necessarily also increase doctrinal rationality.[46] Previously, by inviting juries to make inferences about intention from foresight judges were able, under cover of the *Nedrick* fudge, to enable juries to convict of murder those particularly wicked killers such as terrorists who, while not acting with the intention, direct or oblique, of killing, would not necessarily consider their mission a failure if a death did occur. In short, it brought the *Hyam* test back into the law of murder, but took it out of the substantive law and hid it in the jury room discussion.[47] This no longer seems possible, although it is just possible that *Woollin* is just another step in a masterful juridical 'dialectic'.[48] After all, as explained earlier, *Woollin*, unlike *Hyam*, was a classic case of 'act first think afterwards'. While judges have no business skewing doctrine to convict such killers of murder the same cannot necessarily be said for the likes of Mrs Hyam.

41 Duress of circumstances would not be available here as pregnancy is hardly a life threatening condition; cf though *Bourne* [1939] 1 KB 687.

42 As it did in *Gillick* [1986] AC 112, *Fretwell* (1862) Le & Ca 161, *Salford Health Authority, ex p Janaway* [1988] 2 WLR 442; *Clark* (1984) 80 Cr App Rep 344, for a discussion of the issues see I Dennis 'The Mental Element for Accessories' in P.F. Smith (ed) *Criminal Law: Essays in Honour of JC Smith* (London: Butterworths, 1987); G.R. Sullivan 'Intent Purpose and Complicity' [1988] *Crim LR* 641 and Ian Dennis' reply thereto at 651–3.

43 See discussion at n 20 above.

44 ie cases where the accused does not seek to rely on his good intention.

45 See A.J. Ashworth 'Criminal Liability in a Medical Context: The Treatment of Good Intentions' n 2 above; W. Wilson n 10 above, 282–299; R.F. Schopp *Justification Defences and Just Convictions* (Cambridge: Cambridge University Press, 1998) ch 2.

46 For discussion of this scant considered distinction between rationality and doctrinal certainty see J. Gardner 'Rationality and the Rule of Law in Offences Against the Person' [1994] 14 OJLS 335.

47 cf *Walker and Hayles* (1990) 90 Cr App R 26, and See A Norrie n 14 above generally

48 See N Lacey 'A Clear Concept of Intention': Elusive or Illusory' (1993) 56 MLR 621–3.

© The Modern Law Review Limited 1999

Lord Steyn was sanguine about this consequence, perhaps unsurprisingly in view of the receding risk of terrorism.[49] It seems, then, that in cases such as *Hyam* a judge will fall into error if he gives a direction other than a *Nedrick* direction even, one assumes, if he makes clear that foresight of less than virtual certainty is only one piece of evidence which they must weigh in the scales with the others. Perhaps, however, he reasoned, as Lord Hope also appeared to, that judges were not bound to give the special direction in cases of murderous risk-taking. In such cases, it may be supposed, juries will continue to draw inferences of direct intention from degrees of probability and so moral judgments will still play a part in juries' determinations of definitional fault. In short, *Hyam* will remain part of the law of murder as it is practised[50] if not as it is in theory.[51] An important ramification of this will be considered in the concluding section.

Conclusion

Woollin is not, then, the end of the story. The previous doctrinal ambiguity was made necessary by two separate systemic features. First, sustained judicial reluctance to articulate a workable generalised justificatory defence which would exculpate those whose conduct is morally inconsistent with the criminal intention required by the offence definition. Second, the structure of homicide which, in theory, withholds the label of murderer from those such as terrorists who without aiming to kill still, through the wicked indifference they display, appear to act in violation of the moral injunction 'do not kill'. Since the majority of statements concerning the legal meaning of intention have occurred in the specific context of murder cases, its general meaning has been consistently slewed to support the policy goal that all such 'wicked' killers be labelled as murderers regardless of their actual state of mind. The significance of *Woollin* is that it makes these features of greater practical concern than hitherto. With regard to the first feature *Woollin* may well act as a stimulus for the development of a broadly applicable justificatory defence of good motive if judges, as is quite conceivable, interpret *Woollin* as having incorporated the standard definition of intention in English criminal law. Or, the courts will be forced to further refine the meaning of intention so as to exclude from the scope of criminal intentions those who, with full knowledge but without purpose, bring about an undesired side effect of action taken in support of a legal right or duty. This latter option is likely to have no lesser impact than the former.[52] But without such developments doctrinal rationality will once again be the loser and the juridical 'dialectic' will continue.[53]

With regard to the second feature, doctrinal rationality has already been compromised, if by rationality we mean the process by which the scope of our social obligations are successfully instantiated and articulated in rules. The present crime of murder is both under and over-inclusive. It is over-inclusive by condemning some we might prefer to stigmatise as manslaughterers or not stigmatise at all. It is under-inclusive by finding no place for the terrorist or Mrs Hyam. It is glib to suppose that, as Lord Steyn put it: 'Such cases ought to cause no

49 [1998] 4 All ER 103, 112. Lord Hope was less so, at 114.

50 See N. Lacey 'Philosophy, History and Criminal Law Theory' (1998) 1 *Buffalo Criminal Law Review* 295, generally.

51 A.P Simester and W. Chan 'Intention Thus Far' [1997] Crim LR 704, and see n 27 above.

52 For example, it would change the fault element in accessoryship from knowledge to purpose.

53 See n 48 above.

© The Modern Law Review Limited 1999

substantial difficulty since immediately below murder there is available a verdict of manslaughter which may attract in the discretion of the court a life sentence'.[54] Conviction labels are as important as justice in the distribution of punishment. They identify a precise social obligation dishonoured in the infraction.[55] This is, of course, why we also have various crimes of reckless endangerment[56] to deal with death caused in the course of acting dangerously rather than just manslaughter, as well as crimes of ulterior intent.[57]

Perhaps – is this just the optimism of the rationalist asserting itself? – Lord Steyn has it in mind that the fault element in murder will later be returned to something akin to its pre-1957 position. After all, the House of Lords has shown itself increasingly uncomfortable with implied malice[58] and is wont to give hints on occasions that, when push comes to shove, it will be abandoned.[59] Here Lord Steyn could not resist saying (murder) 'is a species of constructive crime ... This feature of the law of murder may have contributed to the problems which courts have experienced with mens rea in murder. But, unless the House of Lords or Parliament have occasion to revisit this point, the sufficiency of an intent to cause serious harm is the basic assumption upon which any analysis must proceed'.[60] If it is abandoned, and the reference to the 'House of Lords' suggests it may be on the agenda, the obvious replacement is a return to the pre-1957 notion of implied malice. This included (i) an intention to inflict injury so serious as to put the victim's life in peril and (ii) an intention to do an act intrinsically likely to kill but without an intention to cause any physical injury.[61] Such a return would greatly reduce the tendency – to ensure that terrorists and other wicked risk-takers get their just deserts – for judges to play around with the meaning of intention, with beneficial consequences for doctrinal stability and rationality generally.[62]

Unhappily, the only other indication Lord Steyn gave that he was alert to the wider consequences of the decision also indicated he was reconciled to them. In a statement which will send a shiver down the backs of those who esteem general doctrinal consistency and stability as much definitional clarity Lord Steyn said:

> I approach the issues arising on this appeal on the basis that it does not follow that 'intent' necessarily has precisely the same meaning in every context in the criminal law. The focus of the present appeal is the crime of murder.[63]

It may seem then that optimism is somewhat misplaced. The House of Lords in *Woollin* gave no hint that it was committed to restructuring the rules governing criminal liability so as to ensure that those undeserving of the stigma of conviction

54 [1998] 4 All ER 103, 112.

55 In murder that obligation is articulated in the biblical injunction 'Thou shalt not kill' rather than Lord Atkin's watered down version 'Take care not to injure your neighbour'.

56 For example causing death by dangerous driving and arson with intent to endanger life.

57 For a persuasive argument to this effect see J. Horder 'Crimes of Ulterior Intent' in A. Simester and A.T.H. Smith (eds) n 2 above.

58 That is, the intention to cause really serious injury. Their discomfort is due to the fact that having the intention to cause serious injury does not necessarily express contempt for the value of life. It does so only if the intention is to cause life endangering, and not merely serious, injury.

59 See for example Lord Goff 'The Mental Element in Murder' [1988] 104 LQR 40 at 47. Lord Mustill in *A–G's Reference (no 3 of 1994)* [1997] 3 All ER 936, at 945; Lord Steyn in *R* v *Powell* [1997] 4 All ER 545, 551–2 as well as in the instant case.

60 [1998] 4 All ER 103, 107.

61 See Kenny *Outlines of the Criminal Law* (Cambridge: Cambridge University Press, 1962) 153–7; *Russell On Crime* (London: Stevens, 1965) ch 29.

62 Wilson n 10 above, 368–374.

63 [1998] 4 All ER 103, 107–108 For support for a differentiated meaning in cases such as complicity, attempt, and crimes of ulterior intent see G. Williams 'Oblique Intention' [1987] C LJ 417, at 435 et seq. R.A. Duff 'Intentions Legal and Philosophical' (1989) 9 OJLS 76.

© The Modern Law Review Limited 1999

avoid liability through the demands of a system constituted by mutually supportive and internally consistent rules. That is, more simply, that they do so as a result of a legal rather than a political determination. And in the particular case of homicide, it manifested no deeper commitment than to draw a distinct line between murder and manslaughter. The terrain of criminal homicide is beset with greater problems than this. Perhaps such caution was inevitable. Sketching out the framework for a cogent and workable justificatory defence such as might rationally temper the implementation of a universally applicable standard meaning of intention would, on the facts of *Woollin* and in the light of the specified question, be very obiter indeed. It would take a *Gillick* or an *Adams* to justify the wielding of such a broad brush. Similarly, if the House of Lords is interested in returning the fault element in murder to its earlier state, a precondition for its intervention will surely be a case involving a clear-sighted decision to subject another, by violence or otherwise, to the serious risk of death. However, as doctrinal ambiguity decreases so mounts the tension between different parts of the structure of criminal liability and so also increases the pressure for such a development. In this light, the decision in *Woollin* is one small, if equivocal, step in the right direction.

© The Modern Law Review Limited 1999

REVIEW ARTICLE

The Rhetoric of Precedent and Comparative Legal Research

Maurice Adams*

D. Neil MacCormick and Robert S. Summers (eds), **Interpreting Precedents: A Comparative Study**, Aldershot: Dartmouth, 1997, xi + 585pp, hb £65.00.

Interpreting Precedents 'seeks to advance understanding of fundamentals of law and its methodology through systematic comparative and theoretical analysis' (p vii). As its title suggests, the volume focuses in particular on the function, and functioning, of precedent.[1] The bulk of the book is devoted to ten country reports, one chapter for each country studied: Germany, Finland, France, Italy, Norway, Poland, Spain, Sweden, the United Kingdom and the United States. There is also a report on the European Community. It is not clear exactly why the above countries were chosen,[2] but the choice makes sense since it provides a representative cross-section of jurisdictions from the civil and common law. Each of the country reports follows the same structure and seeks to answer the same questions. The volume closes with six chapters offering general jurisprudential conclusions.

The question of convergence

An obvious distinction between the different legal systems studied is that between systems of law in which precedents are recognised as a formally binding source of law (common law) and those in which precedents are not formally recognised as a binding source of law but where they might be followed implicitly or explicitly (civil law). The editors acknowledge that the classifications 'civil law' and 'common law' are unsubtle and, to a degree, misleading in terms of both history and the present (pp 2–4). There does not exist a unified 'civilian' type of legal system, just as there does not exist a single 'common law' type of system. In this study, nevertheless, the typology serves as a starting point, because the authors try to figure out, among other things, what differences might exist between the two systems.

The editors state that although it has sometimes been doubted whether precedents stand for anything at all in civilian legal systems, this study shows that precedents count for a great deal in such systems. They believe, furthermore, that during the latter half of the twentieth century there has been a significant convergence between the two types of system (although important differences still exist) (pp 2, 546). Indeed,

* Faculty of Law, University of Antwerp (UFSIA), Belgium.

1 The volume has a companion on statutory interpretation: D Neil MacCormick and Robert S. Summers (eds), *Interpreting Statutes: A Comparative Study* (Aldershot: Dartmouth Publishing, 1991)

2 The Netherlands for example, possessing aspects of francophone and germanic legal culture, would have been a valuable addition.

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

> precedent now plays a significant part in legal decision making and the development of law in all the countries and legal traditions that we have reviewed. This is so whether or not precedent is officially recognized as formally binding or merely as having other normative force of some degree. For historic reasons, certain legal systems formally discourage or even discountenance the open citation of precedent in judgments at the highest level. But even in these cases, precedent in fact plays a crucial role (pp 531–532).

In the report on Germany, this conclusion is supported by the fact that a very high percentage of the decisions which are published in the official series refer to precedents. It is added, however, that the citations of precedents are self-referential and in most cases contain no detailed discussion (pp 24–25). In the report on Poland, it is said that although judges and lawyers are not formally bound to cite judicial decisions, it is a fact that the courts, since the fall of communism, have shown a greater tendency to quote judicial decisions (p 235). In the report on Italy, we find the view expressed that recent decades have seen a great expansion in the use of precedents. They are now employed as one of the main bases of every judicial decision (p 152). It is added, nevertheless, that conflicts in the case law of the *Corte di Cassazione*, are frequent (p 163). The Constitutional Court in Italy does not feel bound to follow its own precedents, and often changes its mind on legal issues (p 163). The French report demonstrates that although the courts very rarely cite precedents and must not base their decisions on them, precedents are nevertheless important (pp 112–113). In Spain, the case law of the Supreme Court, while usually followed, does not play a major role in day-to-day adjudication. Only certain Supreme Court decisions are deemed authoritative. Of the case law of the Spanish Constitutional Court it is said that it can be interpreted as a strong argument for a given legal solution. In spite of this, the status of precedent in the Spanish legal system is difficult to determine (p 270) and contradictory lines of precedent do exist (p 288). The chapters on the Scandinavian countries are stronger in their proclamations concerning the use of precedent. In Norway the lower courts treat previous decisions of the Supreme Court as determinative of result (p 199), although there is hardly any judicial or academic discussion on this matter (p 214). In Sweden, it is said that the influence of precedent is even greater than in England (p 293). In Finland, precedents have a significant argumentative value in the interpretation of statutes, although a matter cannot, in principle, be solved on the basis of precedent alone (p 82).

What proclamations of this kind show is that there is an explicit awareness of the role of precedent in continental legal systems. But to conclude that the civil and common law are significantly converging is surely an exaggeration. Such a conclusion cannot be based on the type of research undertaken in this book and the kind of information provided by the authors. Of course, everything depends on what the editors had in mind when they wrote that 'it is shown that precedents *count for a great deal* in civilian systems' (p 2, italics added) and that precedent in fact performs a crucial role (p 532). Statements like these suggest more than simply an explicit awareness of the role of precedent. The material question however is not whether precedents play a role, but exactly what kind of role precedents play in the jurisdictions reviewed. I have endeavoured to demonstrate elsewhere[3] that, despite academic arguments to the contrary, Belgian courts mostly cite precedents or case-law, if at all,[4] in an opportunistic manner – ie, when they confirm the

3 Maurice Adams, *De Precedentwerking van Rechterlijke Uitspraken. een rechtstheoretische en rechtsvergelijkende studie* (PhD dissertation: University of Leuven, 1997). My study focuses solely on tort law

4 The highest court in Belgium, the *Cour de Cassation*, never cites any precedent.

© The Modern Law Review Limited 1999

position taken by a court.[5] Precedents are thus mainly used as a device to legitimate a judgment *post factum*. Why should this not be the case in the jurisdictions described in this book? My intuition is that at least on the European continent, with the possible exception of the Scandinavian countries, there is no reason to believe the contrary. In none of the non-common law countries studied do there exist the material conditions necessary for precedent to function effectively. In none of these jurisdictions, for example, is a distinction explicitly made by the courts between precedents that have justificatory force and those which have merely illustrative or supportive value. On the basis of the material presented in this volume, I doubt that it is possible to characterise the role which precedents currently play on the European continent as 'counting for a great deal' or 'crucial'. To prove such assertions, a different approach should have been taken. It is reasonable to believe, in short, that the systems are not converging at all.

Case law plays, and has always played, a certain normative role in most jurisdictions in the civil law world. Every jurisdiction has its own *Donoghue* v *Stevenson*-type cases. In every jurisdiction, courts adjust statutory rules to new social and economic circumstances. If one accepts that judges make law, producing precedents is in the logical order of things. It is a requirement of practical reason that when a judge gives a good reason for a judgement, this notion of a good reason brings with it the concept of a rule or precedent. But the fact that every jurisdiction has its own pivotal cases, or that a functioning system of precedent would be in the logical order of things, does not support the conclusions that precedent is playing a crucial role in civilian systems or that civil law and common law are significantly converging.

In chapter 18, a list of significant remaining differences between common and civil law is presented. In the civil law, there is usually no detailed statement of facts, very little purely substantive reasoning appealing to policy, usually none of the detailed analysis and in-depth discussion of the point and purpose of rulings on issues in prior cases; sometimes, there does not exist a well-functioning official or semi-official system of reporting cases, there is no explicit difference in the rulings between *ratio decidendi* and *obiter dictum*, and no sophisticated methodology of distinguishing precedents otherwise arguably applicable; the lower courts have the liberty to depart from a single higher-court precedent or even a line of precedents, and in five of the countries studied the higher courts consciously and regularly depart from precedent without even mentioning this fact (pp 536–539). According to the editors, these features illustrate that, when compared with the common law, precedent in civil law has a 'somewhat lower normative force' (p 539). This is surely an understatement. Because of all these differences just noted, precedents are unlikely to play a crucial role in the non-common law world (with the possible exception of the Scandinavian countries). There is certainly an explicit awareness of the role of precedent in continental legal systems, and pivotal cases exist and function satisfactorily as precedents, but this volume provides no evidence to substantiate any stronger a conclusion.

5 One of the main differences between England and continental Europe seems to be that, on the continent, there is no presumption that precedents are 'right'. Harris has shown that such a 'presumption of rightness', which he classifies under the heading of 'no new reasons' (meaning that a precedent is in principle right when there are no new reasons to decide the case otherwise), is an important constraining postulate with which the House of Lords works when dealing with the question of whether or not to overrule a precedent. J.W. Harris, 'Towards Principles of Overruling – When Should a Final Court of Appeal Second Guess?' (1991) 10 *Oxford Journal of Legal Studies* 157. This means, among other things, that judgments should be discussed and taken into account before a judgment is delivered and not just *post factum*.

© The Modern Law Review Limited 1999

Methodology

The reason that the editors conclude as they do is explicable by reference to the methodology used. This methodology aims at a rational reconstruction of partial or fragmentary, sometimes internally inconsistent, materials. It calls for a dialectic between assessing the apparent meaning of a part and forming a view of the whole (p 10). This is the position of the well-informed insider who is able (partly) to step out of his legal system and think critically about it. It is a valuable approach within the humanities, of which the discipline of law is certainly a part. The question to be considered here is whether the methodology is suitable for this volume.

The methodology is not directly apparent from the chapters themselves, but emerges from observations contained in the country reports. This is, we might say, an implicit comparative methodology. As Rubin wrote about the companion volume on statutory interpretation: 'It is always possible to derive the [comparative] patterns on one's own, of course, and the book certainly provides the raw material for such observations. But it does not provide us with the individual authors' analysis of those observed patterns, nor does it reveal any additional patterns that the authors themselves might have discerned'.[6] The approach adopted implies that one instinctively knows what is meant when separate authors write that precedent is important (France, p 113) or that only certain decisions of a supreme court are deemed to be 'plainly authoritative' or 'strong argument[s]' (Spain, p 270). But are not these statements too vague to be conclusive? What do they mean? That there is, in general, a crucial role for precedents in the civil law, or just an awareness of them? Can we be sure that the authors of the country reports are describing the actual practices that occur in their legal systems? It would have been valuable if case law were thoroughly analysed, with specific regard to the functioning of precedent, from the lowest level of court hierarchy to the highest: an empirical approach, as it were, akin to that which Paterson used (although he focused on a somewhat different matter) in *The Law Lords*.[7] Such an approach might have provided more reliable information.

Conclusion

The country reports contained in this volume are very informative, but should be approached with circumspection. Many interesting examples of the different sorts of questions concerning the function, and functioning, of precedent can be found in these reports. However, the general comparative methodology employed does not support the conclusion that there has occurred significant convergence between civil law and common law and generates ambiguities concerning the role that precedents play in the jurisdictions studied. To end on a positive note, this volume very clearly illustrates that the concept of precedent is not monolithic. 'Just as the "binding" quality of precedents is subject to a variety of interpretations, so may there be more than one way of using past decisions to guide the present and the future' (p 497). For all that the tone of this essay has been primarily critical, it ought to be emphasised that a valuable lesson to be drawn from *Interpreting Precedents* is that the formal binding quality of a precedent is but one element that is relevant to understanding what kind of authority ought to be attributed to a judicial decision.

6 E. Rubin, 'Review of "Interpreting Statutes. A Comparative Study"' (1993) 41 *The American Journal of Comparative Law* 135 at 135

7 Alan Paterson, *The Law Lords* (London: MacMillan, 1982).

© The Modern Law Review Limited 1999

REVIEWS

Marjorie B. Zucker and Howard D. Zucker (eds), **Medical Futility and the Evaluation of Life-Sustaining Interventions**, Cambridge: Cambridge University Press, 1997, 201pp, hb £45.00, pb £16.95.

Questions concerning the use of end of life technology continue to dominate both the political and ethical landscape. In the UK, a recent consultation document issued by Lord Chancellor's Department on decision-making for the mentally incapacitated, charged with looking at the issue of advance directives afresh, states: 'Advances in health care mean not only that people are likely to live longer, but also in circumstances and with illnesses and disabilities which previously would have led to the person's death'. The implications of this are profound. As a greater proportion of the population survives into old age, society must ask whether it is right to use medical technology to maintain frail lives at all costs, and policy-makers must devise the legal and ethical foundations upon which decisions may be made for patients who can no longer consent to, or refuse, invasive life-prolonging interventions. Harder still, we can no longer shirk from addressing the unpalatable issue of whether it is morally justifiable to use a significant proportion of scarce resources to provide expensive end-of-life therapies for a relatively small number of patients, particularly where those interventions offer little or no medical benefit to the patient.

Medical Futility expertly addresses all of these matters, as well as surveying the clinical, ethical, legal, religious and personal dimensions involved in end-of-life decision-making. Defining what is meant by 'futility' is clearly a central theme throughout the various chapters, although a consensus remains elusive. What emerges, though, is that futility is rarely defined in absolute terms; rather, its meaning has shifted to include treatments the burdens of which do not justify the benefits, particularly those interventions which are capable only of maintaining a patient's exceedingly poor quality of life. Decisions to withhold life-sustaining therapies may, nonetheless, appear to conflict with health practitioners' ethical and legal duty to preserve life. This book considers the extent to which withholding treatment is strictly a technical medical decision, and challenges previously held notions that physicians should be the sole or primary decision-makers in such situations.

A strength of this eminently readable book is that the contributors are drawn from a wide range of disciplines. To highlight the personal ramifications of the subject matter, the editors have included a most moving personal testament, 'Death with Dignity?', by the widow of the late Paul Brophy, subject of a landmark US end-of-life legal case. Furthermore, it examines the question of futility from a number of different angles, including chapters on futility from the perspective of paediatrics, the nursing home environment, and general practice, as well as the more frequently discussed critical care setting. Thus, for example, Frader and Watchko point out that in a paediatric context, the futility debate is shaped both by the deep-seated cultural belief that children should not die before some imaginary natural life-span, and the inherent medical uncertainty in predicting eventual medical outcomes in this area.

The contributors are all based in the United States and the issues are represented primarily from a US perspective. For the main part, the ethical arguments are equally relevant to the situation in the UK, although there are a number of differences which

© The Modern Law Review Limited 1999 (MLR 62 3, May) Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 350 Main Street, Malden, MA 02148, USA

colour the American experience. One contrast is the way in which health care is funded. In the US, the principle of managed care embraces a requirement to reduce health care expenditure by eliminating non-beneficial interventions, and this is clearly influential in the futility debate. It would be naïve, however, to pretend that resource implications do not have some part to play in futility discussions in our own cash-constrained NHS. However, in terms of treatment-limiting protocols, both here and in the US, there is an unwillingness to address the resource issue head on. Rather, the front is maintained that each individual end-of-life decision is a clinical decision to be taken by physicians on the individual facts of the case. Another interesting difference is the reliance in the US on clinical ethical committees to assist in end-of-life decisions. This raises the issues of institutional self-interest, for as Frader and Watchko point out, one single paediatric case may represent a significant economic burden for a particular hospital.

Advance directives have been seen as a useful device to limit unwanted interventions, made by people whist still competent, to apply at some future point when they become mentally incapacitated. As Bartoldus points out in her chapter on the nursing home perspective, in the US, the 1991 Patient Self-Determination Act makes it a legal requirement for nursing homes (and indeed all federally funded institutions) to discuss advance directives with residents upon admission. Were such a policy to be implemented here, clinicians might at least be able to identify those patients who did not want to receive unwanted life-sustaining interventions, but of course this still requires policies to be devised for the remainder of patients who have not indicated a preference.

The scope and validity of advance directives has been addressed by courts both here and in the US. But is recourse to law a helpful way of solving the futility debate? Particularly, can the law determine whether treatment should be given which the patient or the patient's family wish to be initiated or continued, but which the doctor considers futile? Prip and Moretti's chapter highlights the difficulty in attempting to derive from the principle of respect for autonomy not only a legal right to refuse treatment but also a legal right to receive treatment. Particularly in the US, the negative right of freedom from interference has a long legal tradition and is rooted in the constitutional right to liberty, and the common law right to be left alone. In the US, as in the UK, courts have been reluctant to confer individual rights of access to health resources. In *Gilgunn* v *Massachusetts General Hospital* (1995), a jury held that doctors were not negligent in their decision to withdraw ventilation and issue a Do Not Resuscitate (DNR) order despite the patient's daughter's objections, on the rationale that the intervention was futile. The judiciary, Prip and Moretti conclude, offer poor tools for reaching consensus in this area.

The cost of providing futile treatments is a recurrent issue. In his introduction to the book, the renowned US ethicist, Alexander Morgan Capron, points to the current re-examination of the principle of autonomy and asks whether there are not times when the claims of the community outweigh individual preferences. This sentiment is picked up by Murphy in his chapter on the economics of futile interventions. He argues that although truly futile interventions are rare, '[w]e need to make the transition – painfully and slowly – from the rugged individualism that formed our country to a more communitarian ethic that will help us survive (and perhaps thrive) in the next century'. Arguing persuasively about the 'death of common sense', he suggests that although the cost-savings in this area may be small, if society cannot deal responsibly with futile interventions – by which he connotes the merging of good ethics, good patient care and good economics – then

© The Modern Law Review Limited 1999

there is little hope of dealing responsibly with the much larger question of inappropriate intensive care.

This is a fascinating and thought-provoking collection of essays, which will provide stimulating reading for health care professionals, students, lawyers and ethicists. Ultimately, though, this book offers the message that discussions about futility and end-of-life need to involve all members of society, and that this must become a collective, rather than a professional, debate.

*Julie Stone**

Michael Tonry and Kathleen Hatlestad (eds), **Sentencing Reform in Overcrowded Times**, New York : Oxford University Press, 1997, xv + 288 pp, pb £17.50, hb £37.50.

As the editors are quick to point out, the title of this collection of short articles is in reality a play on words since all fifty-seven contributions first appeared in the American journal, *Overcrowded Times,* which specialises in sentencing and corrections issues. This is the second such endeavour, the first, *Intermediate Sanctions in Overcrowded Times*, having been published in 1995. Although both books boast a distinguished list of contributors (in the text under review the section on England in chapter 3 contains pieces by Ashworth, Wasik, Moxon and Morgan) I found it difficult to reconcile the fact that the articles are summaries, written in 'nontechnical language for a nonspecialist audience', with the declared aim 'to make reliable, authoritative information on important research and program developments available to policymakers, practitioners and researchers' (p v). It is precisely their lack of detailed contextual analysis which renders these vignettes of limited utility, particularly to policymakers and researchers.

It is certainly true that there is, with notable exceptions, a paucity of detailed comparative material in the sentencing field. Virtually all the relevant texts share a weakness in common with the present volume in failing to recognise the need to integrate comparative material within a theoretical framework which, at the very least, acknowledges the different jurisprudential traditions of numerous legal systems (for example, between Roman and common law systems). In chapter 1 of their volume, Tonry and Hatlestad fail to work through the implications of what comparative research can actually achieve, since they fail to address major theoretical questions concerning the nature of law, justice and normative behaviour within diverse jurisprudential traditions. At the very least, this requires a detailed analysis of 'history' and 'process' (p 5); but the contention by the editors that 'there is virtually no scholarly or professional literature on sentencing policy changes' is certainly not borne out by the English literature (see, eg Windlesham, *Responses to Crime: Volume 3, Legislating with the Tide* (1996)).

The remainder of the introductory chapter provides thumbnail sketches of sentencing policy developments in both the United States and elsewhere which, although valuable as an overview, are more adequately provided in the case of the United States by Tonry himself in *Sentencing Matters* (1996). Of particular interest to the English reader are the twenty-four contributions in chapter 2 outlining developments in some thirteen American States and nationally. However, chapter 3, which deals with sentencing reform outside the United States, is more problematic. Eight of the nineteen contributions are devoted to developments in

* University of Birmingham Medical School.

© The Modern Law Review Limited 1999

England and Australia which occurred in the early 1990s and hence do not assess the genesis and nature of recent governmental sentencing policy changes. For instance, the English essays chart developments surrounding the Criminal Justice Act 1991 and its immediate aftermath, whereas analysis of the more recent phenomenon of 'populist punitiveness' and the gradual dismantling of any principled approach to sentencing would be of greater utility to policymakers. Nevertheless, direct lessons can be learnt from detailed comparative material on contemporary developments in similar sentencing systems. An example is provided by the differing impacts of 'truth in sentencing' regimes introduced into Victoria and New South Wales. Apart from England and Australia, descriptions of reforms elsewhere are limited to one or two contributions and, with the notable exceptions of South Africa and Switzerland, many have received more extensive and directly useful coverage elsewhere in other contexts. In his introductory essay to chapter 3, Kuhn tackles the daunting task of comparing prison population trends in seven European countries. Quite apart from the fact that it is notoriously difficult to discern meaningful relationships between sentencing policy and prison populations within a domestic (let alone an international context) Kuhn recognises that:

> It is difficult to make international comparisons of corrections practices, partly because cultural ideas about crime seriousness and punishment severity are effected by social, economic and political conditions, and these vary between countries. Such comparisons are made more difficult by the diversity of criminal justice systems and repressive institutions (p 125).

Many of Kuhn's conclusions are, nevertheless, disappointingly simplistic. For example, he suggests that the French experience with amnesties and other stays of sentences temporarily mask upward trends in prison populations, but do not mask prison overcrowding (p 129). Similarly, when commenting on Austria, Kuhn observes that, as in Germany, the abolition of short-term imprisonment does not seem to have been a long-term solution to prison overcrowding. Such general observations simply reinforce the need for detailed contextual analysis of the emergence of specific penal measures.

In my opinion the most instructive chapters for UK policymakers and sentencing scholars are 4 and 6, which focus, respectively, on race and public opinion issues in sentencing. Although cultural context is clearly vital in transposing or evaluating racial and public opinion issues, Tonry and Hatlestad make an excellent point in their introductory note to chapter 6 which echoes a common perception of recent UK sentencing policy and contemporary analysis of past penal strategies: 'The key question is whether public attitudes became harsher, and public officials followed, or whether conservative politicians eventually used crime as an issue to frighten Americans and then offered to respond to those fears by enacting and carrying out harsh policies' (p 249).

In conclusion, the limitations of this collection do not detract from its fundamental value as a quick reference guide or snapshot of sentencing reform in the United States and elsewhere. But the work singularly fails to integrate and assimilate comparative accounts. In view of the political, cultural, sociological and economic forces that shape contemporary sentencing policy in Western industrialised democracies in particular, I would suggest that multi-disciplinary, theoretically rigorous analysis of contemporary sentencing policy is now well overdue. Regrettably, since this has yet to be attempted in one single country, a comparative project along such lines is likely to be a long way off.

*Ralph Henham**

* Nottingham Law School, Nottingham Trent University

© The Modern Law Review Limited 1999

Marc Maresceau (ed), **Enlarging the European Union – Relations between the EU and Central and Eastern Europe**, Essex: Longman, 1997, xxv +403pp, hb £40.00.

Relations between the EU and the Central and East European states (CEES) have been one of the pivotal forces in the development of internal and external EC and EU policy especially in the last ten years. The so-called 'first generation agreements' signed in 1988 with Hungary, Poland, the Soviet Union, Bulgaria and Romania were quickly followed by more detailed and extensive association agreements, the Europe Agreements. Currently ten states have signed Europe Agreements: Hungary, Poland, the Czech and Slovak Republics, Bulgaria, Romania, Estonia, Latvia and Lithuania and Slovenia. States further to the east of the EU have not been offered Europe Agreements. Instead a weaker form of agreement, the Partnership and Cooperation Agreement, has been signed with Russia, the Ukraine, Moldova and Belarus. Many of the CEES had long made explicit their intention of joining the EU. The Commission and Council response has been piece-meal. Guidance on accession criteria has been developed through the Council Summits and the Commission published a White Paper providing guidance on harmonising legislation to the *acquis communaitaure* related to the Single Market. By 1996 all CEES with Europe Agreements had applied to join the EU.

As we now enter what the Commission refers to as the 'pre-accession phase' it is clear that the importance of these relations will continue to dominate and dictate the pace of policy development. This book provides a welcome addition to the recent growth of academic literature in this area for two main reasons. First, it provides a very useful analysis of the current state of legal and non-legal relations between the EU and the CEES. Secondly, it attempts to highlight some of the mains themes which are likely to shape the enlargement debate.

The book is divided into three parts. Part I provides an overview of the main developments. Part II focuses on the Europe Agreements and pre-accession strategies. Part III examines the issue of enlargement considering the impact upon the economies of the EU and CEES, Common Foreign and Security Policy and globalisation. The diverse list of contributors is one of the book's principal attractions; contributors are drawn from academia, government, the Commission and the civil service and from both within and outwith the EU. Many have non-legal backgrounds.

One of the book's strengths is that its contributors clearly explain the complex relationship between the legally binding instruments, such as the Europe Agreements, and non-legally binding instruments, such as the White Paper. Muller-Graff's paper in particular presents an interesting analysis of the developments in Commission policy. The book also provides a very useful examination of the interaction of EC legal instruments (here the Europe Agreements) with broader international trade agreements (in this case the WTO Agreement and the TRIPS Agreement). In addition it provides a detailed analysis of specific areas such as competition law, movement of persons and safeguard measures.

A number of contributors criticise the EC response to the CEES. This criticism is perhaps a little unfair. If the Commission response was less than helpful it was largely due to Commission surprise at the pace of events within the CEES. It took some time before a dedicated section in the shape of DG1A was established to steer EC activitity in a more coordinated way.

Much of the analysis follows the traditional style of legal academic study in the field of European external relations. Contributors largely present a textual analysis

© The Modern Law Review Limited 1999

contrasting the scope and content of the Europe Agreements with the White Paper, the European Economic Area Agreement, the WTO Agreement or the EC treaty. Such research is crucial to an understanding of the complex relations between the EU and the CEES but presents only a partial picture of those relations. Very few contributors attempt to consider the impact or implementation of the existing legal and non-legal instruments within the CEES. Chapters 6 and 7 examine attempts to implement competition law and state aids provisions. In both cases the writers concentrate upon the measures adopted within the framework of the Association Council. Whilst the Association Councils do have an important role to play in the implementation process they are not the only actors within that process. An approach which focuses solely upon the implementation mechanisms provided within the Europe Agreement ignores the importance of implementation mechanisms within the CEES. An understanding of the systems of public law within these states will become indispensable for academic and non-academic commentators alike in attempts to monitor progress towards accession in the CEES.

The limitations of an analytical approach which is purely text based may be seen again in chapter 8 which examines dispute settlement within the Europe Agreement. The chapter provides a solid critique of the limitations and inherent difficulties of the arbitration procedure provided in the Europe Agreement. It fails however to consider the political importance of the arbitration procedure. For example, when the arbitration procedure was initiated under the Polish Europe Agreement, it was not necessary for the process to go any further since the threat of arbitration provided the impetus for the Parties to reach a solution. In addition, there have been many other disputes since the entry into force of the Europe Agreements which have been resolved using diplomatic and political means.

One of the most interesting contributions is by Jacke Saryusz-Wolski, formerly the Polish Under-Secretary of State for European Integration and Foreign Assistance. He was a key figure in the negotiation and implementation of the Polish Europe Agreement and in the Polish application for EU membership. Saryusz-Wolski argues that the debate surrounding the costs of enlargement to the East is fundamentally flawed since it has been largely based upon calculations which ignored many social and economic factors relevant to the CEES. He stresses that the budgetary expenditures caused by enlargement cannot be considered separately from the economic and political effects of integration. His comments are not merely relevant to the study of relations with the CEES but would apply equally to any academic work in the field of EU external relations. He cautions against studies 'which ignore ... socio-economic conditions' or 'which taken into account only selected factors'. It is imperative that future EU external relations research attempts to pursue this type of inclusive approach.

Relations between the EU and the CEES continues to develop apace. Since the publication of this book the Commission has published Agenda 2000 which recommends that the Czech Republic, Estonia, Hungary, Poland and Slovenia join Cyprus in the next wave of accessions to the EU. The Council has accepted these proposals and negotiations for the first group of CEES were due to begin on 1 April 1998. This book provides an important contribution to the academic debate which must take place as the process of enlargement unfolds.

*Lynn E. Ramsey**

* School of Law, University of Bristol.

© The Modern Law Review Limited 1999

Hugh Corder and Tiyanjana Maluwa (eds), **Administrative Justice in Southern Africa**, Department of Public Law, University of Cape Town, 1997, vii + 206 pp.

This collection of papers was delivered at a workshop at the University of Cape Town in March 1996, the latest in a series of conferences hammering out the requirements that need to be addressed by an enacted code on administrative justice. The right to lawful administrative action is entrenched in the South African Interim Constitution, but the Final Constitution goes futher in requring the legislative enactment of administrative justice. It is this project that forms the centrepiece for these papers.

The legislative project is not unproblematic. The right to administrative justice has to be balanced against the mandate of the ANC government to deliver reforms under its reconstruction and development programme, an exercise which involves considerable expansion of the regulatory powers of the public sector. In light of this, the conference organisers obviously felt that governmental power should not be the only focus of their concerns and invited contributions exploring and questioning the scope of public power itself.

Craig deals with the analytical difficulties of the source/function divide, with excessive formalism on the one hand and uncertainty on the other. We tend to assume in this country that these problems are engendered by Section 31 of the Supreme Court Act 1981, but the scope of public power is clearly a global issue. In South Africa, the question arises in the form of vertical or horizontal application of constitutional rights; and the answer to that question, like most solutions to the public/private dilemma, is a fudge. Sometimes it is appropriate that private parties should be bound by some sort of higher order law, sometimes it is not. For this reason, the public/private dividing line may not be as impermeable as much academic debate suggests. Davis points out that the requirements of public law are filtering down into private law anyway; contracting parties are obliged to observe the doctrine of *bona fides*, and the principles of procedural fairness and legitimate expectation play a central role in employment relations. In Davis' view, the restrictive scope of traditional public law discourse no longer reflects the true picture of society. It is the economic structure of society which creates the vertical relationship, not the presence of a governmental element or public function or whatever distinguishing feature is relied upon to keep public law away from private relationships.

The participants from neighbouring countries have a bleak picture to paint of parliamentary controls on the power of the executive. Giving the Kenyan perspective on parliamentary scrutiny of government action, Ojwang observes that few of the controls in the Westminster prototype can be said to exist other than in theory; select committee reports go unimplemented, private members' motions never reach the statute book, and the no confidence vote, which lies at the root of parliamentary control over government power, has been been rendered nugatory by the whittling down of opposition power. There is little fertile ground for the development of strong opposition parties in African societies, where there are no established power blocs to represent, only ethnic groups with what Ojwang calls a 'mixed bag' of hopes and aspirations. No amount of tinkering with the technical structure of constitutions can change this picture. Ojwang's forceful message goes to the heart of the administrative justice problem. If the executive is not permitted to address the country's development problems with as few restraints as possible, largely agricultural and rural societies may never reach the level of economic development in which these constitutional models and modes of accountability

© The Modern Law Review Limited 1999

have any real meaning. However, Ncube's acount of the Zimbabwean experience shows that unfettered governmental power is no solution. On the contrary, the Zimbabwean constitution, far from being ignored, has been so structured as to legitimise unregulated presidential power, and any inconvenient interpretations of the individual provisions provided by the courts have been rapidly followed by parliamentary amendment. In general, however, control of public power by the judicial branch has been shown to be more effective in the African context than parliamentary restraint. But this may amount to nothing more than the odd wrist-slap, triggering a raft of legislative powers buttressed by ouster clauses; the reason for the (temporarily) more effective controls imposed by the courts may have less to do with the perceived importance of the independence of the judiciary than the fact that judges come from a different socio-economic background to the rulers, a factor which, in Nsereko's view, judges forget at their peril, particularly when invalidating administrative decisions on grounds of unreasonableness.

There is nothing new in this tension between the demands of welfare reform and the need to control public power. The architects of the New Deal in the United States strove to address similar problems, zealously removing obstacles to social change by limiting the power of the judiciary to protect vested interests in property and liberty and by delegating power to regulatory agencies. Sunstein observes that the Americal Administrative Procedure Act was borne of the compromise between the New Deal enthusiasts and their opponents, so that these broad delegations of power would be permitted on condition that they were accompanied by procedural safeguards, monitored by the judiciary. The lessons this holds for South Africa, however, are limited. If the horizontal application of some of the rights in the Final Constitution is anything to go by, the legislative enactment of administrative justice will not confine itself, as the APA does, to statutory bodies. But there might be something to be said for adopting the American courts' extension of judicial review to bureaucratic *inaction* as well as positive action, since it enables the courts on occasion to hold the government to its welfare promises, a potent legitimising role for judicial review.

This is crucial to the administrative justice project. Jowell's paper sets out the developing principles of judicial review in England and explores the influence of Community law. He observes that whilst judicial deference is now a thing of the past, courts are 'sensitive to their own limitations ... Where social and economic policy and resource allocation is involved, the courts give a wide margin of appreciation to the decision makers, and some decisions are not considered justiciable at all'. But those in charge of the purse strings in South Africa would not take much comfort from the recent case law in this country. For every *ex parte B* [1995] 2 All ER 129, there is an *ex parte JCWI* [1996] 4 All ER 385, or an *ex parte Witham* [1997] 2 All ER 779. Judges cannot be relied upon to stay their hand, particularly if social and economic claims are packaged as appeals to justice.

The right to administrative justice is forcing South African lawyers to open their eyes to non-legal forms of redress. Controlling public power is not just about restricting the excesses of government. More misery is caused by the helplessness of ordinary people before the deadweight of bureacratic inefficiency than any amount of executive megalomania. Plasket describes one particularly poignant incident in his role as director of the Legal Resources Centre in Grahamstown. An old man had had his disability benefit stopped because the computer records showed that he was dead. No amount of medical testimony to the contrary would persuade the Department of Social Services to budge. If the Public Protector or Ombudsman had been around at the time, says Plasket, he would have availed

© The Modern Law Review Limited 1999

himself of his offices. In order to be an effective avenue for remedying administrative inefficiency, recourse to the Public Protector would have to be legally aided. The Public Protector himself, Selby Baqwa, makes no mention of this in his paper. Both Plasket and Baqwa are curiously silent on an issue which is central to most critical debates about the Ombudsman in this country: his powers of enforcement. The offices of the Human Rights Commission are also open to anyone complaining about a violation of their constitutional rights, although, as Govender points out, the fiscal restraints on the Commission will make it more sympathetic to systematic abuses than to individual applications.

Birkinshaw's paper on freedom of information and the powers of the Ombudsman describe in detail the position in UK law at the time the workshop took place. Most of what he has to say has now been superseded by the publication of the White Paper on Freedom of Information. There is however an interesting twist in the developments subsequent to the holding of this workshop. One of Birkinshaw's most pressing concerns is the introduction of some sort of legislative protection for whistleblowers. The White Paper does not concern itself with this issue at all, apart from a passing reference to the latest Private Bill being debated in Parliament. The draft Open Democracy Act in South Africa, on the other hand, contains detailed provisions excluding whistleblowers from civil or criminal liability. As with many of its recent legislative initiatives, South Africa appears to be at least one step ahead of us.

The papers collected in this volume contain much of interest to academics and practitioners in the United Kingdom. The issues of administrative justice have a sharp focus in the South African setting, with the great progressive experiment in constitutionalism and human rights on the one side, and the looming dangers of single party domination on the other.

*Rosalind English**

Michael Taggart (ed), **The Province of Administrative Law**, Oxford: Hart Publishing, 1997, xxviii + 381 pp, pb £20.00, hb £45.00.

Public lawyers have long been concerned with arguments about whether it is possible to set any clear boundary between public law and private law – and, if so, where it should be drawn. There are two reasons for this concern. The first, practical reason is that bodies of a public law nature are generally – when acting in a public law capacity – subject to special standards of review which are not (or are not officially) applicable to private bodies. Some legal systems – including English law – also accord special procedural protections to public bodies in cases of an exclusively (or sometimes of a predominantly) public law nature. The second reason is rooted in considerations of philosophical and constitutional legitimacy, in that any substantive distinction (and most procedural distinctions) between public and private law must presuppose that public and private law actors have distinct roles and statuses. In terms of justice and political morality, this has important implications – which will vary according to one's background political and constitutional theory – for the role which one feels the state should serve by contrast with the citizen, for the basis employed to determine whether the

* Merton College, Oxford.

© The Modern Law Review Limited 1999

government is acting legitimately, and for the constitutional justification for the courts exercising a supervisory jurisdiction via judicial review.

The essays collected in *The Province of Administrative Law* seek to reconsider these issues in the light, first, of recent attempts to reduce the role of the state in western, liberal democracies – via the application of market standards to the performance of government and the privatisation, deregulation and contracting-out of previously governmental functions – and, secondly, of processes of globalisation, which arguably render the very concept of the nation state superfluous. The essays are not formally interlinked, but two recurring questions occur. The first is whether public lawyers should respond to these recent trends by abandoning or merely refashioning the public law-private law divide. The second – more traditional – question is whether it is possible, even without the recent trends, for an adequate public law-private law divide to be drawn; and, if so, where it should fall. In English law in particular, both questions are underpinned by the ongoing debate about whether judicial review should be seen as resting on the orthodox *ultra vires* model, or whether some newer model such as control of public power should be adopted. A related question which arises in some essays is how interventionist the courts should be in scrutinising administrative action. It must be stated at the outset that *The Province of Administrative Law* contains some important and thought-provoking essays which are likely to provide the starting point for much future discussion about this area. It is for this reason that some of the arguments are subjected to critical scrutiny in the following paragraphs.

By way of introduction to the first question, Taggart asserts that the 'profound changes brought about by deregulation, commercialisation, corporatisation, public sector downsizing, privatisation and globalisation have fundamentally altered the political and social landscape' within which administrative law must operate (p 2). The exact meanings of privatisation and outsourcing are explored in chapter 4, by Aronson. Their significance for administrative law is, however, neatly captured in Hunt's observation, in chapter 2, that they have resulted in the state being 'reconceived on the model of market ordering' (p 21), and in Aman's suggestion, in chapter 5, that '[a]lmost all of these reforms are market-oriented ... they either substitute markets and the private sector for regulatory regimes or have public agencies use market approaches, structures and incentives to achieve their regulatory goals' (p 90). Aman argues that in the USA, the range of ostensibly public functions deemed capable of being fulfilled in the private sector has been enlarged, particularly via the privatisation of 'operational' (as opposed to policy-making) aspects of government work; and within the public sector itself, government agencies have been re-cast along the lines of federal corporations, greater reliance has been placed on bargaining and negotiation in the exercise of agency discretion and government regulation, and market-oriented cost-benefit analysis tends to be used in place of public interest considerations in decision-making. These reforms can be tied in, Aman argues, with broader processes of globalisation – decentralisation and efficiency have been prioritised in both public and private sectors owing to the emergence of global technologies and worldwide markets. There is now an unprecedented degree of interconnectedness among national economies – which, together with intensive inter-company competition at global level, leaves a much reduced role for national government as traditionally understood. Aman thus argues that '[i]n the global era, administrative law now appears to be moving away from its role as legitimating new extensions of public power, to legitimating new blends of public and private power and/or private power used for public interest ends' (p 95).

© The Modern Law Review Limited 1999

As Hunt observes in chapter 2, some commentators believe that this reconception of the state may have created a dangerous accountability vacuum in English law. If the legal basis for the provision of many formerly governmental services has shifted from public law to contract, the provision of those services may have been placed beyond the reach of effective constitutional control (whether legal or political), while an inconsistency may have developed between the new 'commercial' values of the public sector and the more traditional public law values espoused by the courts in judicial review. Hunt is not concerned to debate the merits or demerits of this reconception of government. Rather, he aims to show that 'the contractualisation of government only presents a threat to the province of public law so long as English courts remain in thrall to the Diceyan version of constitutionalism' (p 38). Whether the courts are still in thrall to such a version – based on balancing respect for Parliamentary Sovereignty against essentially private rights – is, he suggests, currently uncertain. Courts have begun to develop a more modern account of constitutionalism – overtly concerned with safeguarding constitutional liberties – but their progress has been mixed. The modern account will be safe, Hunt argues, if courts accept that they can assert a public law jurisdiction over powerful bodies *regardless* of the source of their authority.

This argument – which might, in Loughlin's terminology, be categorised as liberal normativist in nature – adds important weight to the case for recognising a non-*ultra vires* model as the basis for judicial review, as well as providing a way of meeting fears about an accountability vacuum. Hunt's account of the emergence of a new form of constitutionalism cannot be based solely on the cases he cites concerning the public law-private law divide, however. For, given the emphasis Hunt places on the protection of constitutional liberties, his account must also show that the courts are becoming less deferential towards the executive and are instead insisting on fully protecting such liberties in cases concerning the *grounds* of judicial review. Admittedly, Hunt has discussed this area at length in his book, *Using Human Rights Law in English Courts*, but it would have been helpful to refer to such discussion in chapter 2. This is a minor criticism of an otherwise valuable and stimulating contribution, however.

Mullan appears slightly more radical in his aim in chapter 7, in that his argument seeks to abandon the orthodox public law-private law divide. Mullan suggests that what have traditionally been seen as core public law principles are now being applied outside the conventional public law sphere in English and Canadian case law. This development must be viewed in the light of the deregulation and contracting out processes, and an assimilation of public and private may be desirable – at least in the areas affected by these processes – given the imperatives of globalisation. Mullan argues that while Canadian courts are increasingly applying 'public law' rules of natural justice in the context of contractual tendering, together with related notions of good faith in the performance of contractual obligations, there are some areas in which public law standards should continue to be inapplicable (religious orders are cited as an example). Nevertheless, a rigid public law-private law divide should be eschewed. The courts should begin by asking whether the applicant is credibly asserting that a wrong has been committed, before asking whether there is anything about the status of the body which committed that wrong which calls for a special standard of review. Since the common law seeks to control the exercise of power in its various manifestations, it 'will at least in some measure provide checking and accountability mechanisms in those areas where the state in accordance with current trends goes about deregulating, privatising and corporatising' (p 159).

© The Modern Law Review Limited 1999

Mullan's argument resembles Hunt's in that it seeks to deal with the accountability vacuum problem and must imply an abandonment of the *ultra vires* model as the basis for judicial review. Mullan may well go beyond Hunt, however, in terms of the range of bodies which he seems willing to see subjected to the type of controls currently associated with public law. Unfortunately, this leaves Mullan's argument facing a difficulty – largely unaddressed by its author. For Mullan's suggestion that public law-private law divide issues should be considered only after the existence of a wrong has been established may turn out to be not so different, in practice, from existing judicial approaches – at least in areas which have not been directly affected by privatisation and contracting-out. For while the distinction would not, as at present, determine the *definition* of the wrong in question, Mullan seems to envisage that it would affect the *intensity* of judicial scrutiny and the nature of the *relief* granted. But even if Mullan's approach falls short of its stated aims and is not as radical as its author implies, it may still constitute a usefully pragmatic way of resolving some of the current difficulties.

While Hunt and Mullan would in practice be happy to retain – to differing degrees – distinct realms of public and private law, MacLauchlan advocates an entirely new legal approach to the reformed structures of government in chapter 6. After analysing the concepts of negotiated rule-making now popular in the public sector – which he labels New Public Management techniques based on the notion of responsiveness – MacLauchlan argues that traditional public law, associated with the values of due process, rules, order and coherence, should be replaced with a new conception of public service law. Public service law would seek to strike a balance between upholding the rule of law and public service professionalism on the one hand, while being responsive to the political priorities of government and its 'clients' on the other. It would not be concerned with the traditional public law-private law divide, seeking instead to uphold public interest values and perspectives whoever is performing the function under scrutiny. Under this approach, law should respond to the incentives and performance measures typical of New Public Management, rather than fostering a command-sanction, rules-based view of government which is likely to stifle private initiative. This provides an interesting contrast to the court-centred approach advocated by Hunt, but one uncertainty in MacLauchlan's argument is that his conception of the rule of law is never made entirely explicit. It is, in particular, unclear how this conception would relate to what MacLauchlan describes as the pragmatic central question of public service law, namely 'who is the client?' (p 133). Until this relationship is clarified, it seems possible that public service law might collapse into a market-oriented version of the earlier functionalist theories from which MacLauchlan seeks to distance himself, quite apart from failing to explain exactly how such a regime would make government accountable to its citizens in practice.

A similar diversity of views about whether the public law-private law divide should be abandoned – and how it should operate if it is to remain – can be found in the essays which consider more traditional aspects of the distinction. Craig's argument, in chapter 10, presupposes the continued existence of the distinction, but subjects the bases on which it might be drawn to critical analysis. Craig suggests that the *ultra vires* model and its main competitor – the notion that judicial review exists to control the exercise of public power – are both examples of what he terms an institutional approach, based on the idea that the amenability of a given body to judicial review will depend upon how closely it resembles a paradigm, purely 'public' body, however defined. *Ultra vires* is as much an institutional approach as control of power since it assumes that the existence of a statutory source for a

© The Modern Law Review Limited 1999

body's authority identifies that body as being of a 'public' nature. Craig rehearses the weak points of both models – focusing on a body's source of authority can be over- as well as under-inclusive, while control of public power can be imprecise and simply begs the question 'what is public?' Institutional approaches can be complemented, Craig argues, by a more pragmatic functional approach, based upon an examination of the consequences of ascribing the label 'public' or 'private' to a given body. Thus, a body would be labelled 'public' if it was felt (within the constraints imposed by the structure of the constitutional system in question) that it should be subjected to distinct procedural and substantive rules. Whether 'public law' rules should be applied would thus depend upon the *consequences* of applying them, as opposed to the *possibility* of such consequences being made to depend upon whether 'public law' rules were deemed to apply to the body concerned.

Craig's account of the institutional approaches involves a useful debate with Allison. In chapter 4, Allison seeks to build on an argument originally developed in his monograph, *A Continental Distinction in the Common Law*, namely that a clear conception of the state is necessary for an adequate public law-private law distinction to work. Allison argues that such a divide cannot work when English law contains no clear concept of the state. Public lawyers instead rely on or imply three separate tests when seeking to distinguish between bodies and decisions which fall within and outside the scope of judicial review. The first asks whether the body concerned is performing a public or 'governmental' function. Rather than trying to describe the state as such, the second is contextual: it suggests that the terms 'public' and 'private' are appropriately attached to bodies where they are specified according to some predetermined set of norms. The third employs concepts of public and private power to determine the capacity in which a body is acting. Allison suggests that all three tests are question-begging, in that they use concepts of public/governmental and private without supplying adequate criteria for distinguishing between them. Allison concludes that English public lawyers would do better to come clean and admit that they need to find a definition of the state in order adequately to separate public from private law.

Craig, by contrast, suggests that it is doubtful whether a definition of the state would make the public law-private law divide any clearer. First, the fact that French law employs a concept of the state does not necessarily make it any easier to tell which bodies fall within the state's remit, given that its boundaries remain contestable. Secondly, using a concept of the state has not helped to determine which bodies should be the subject of any relevant rule of public law. Thirdly, EC law suggests that the concept of the state can vary from case to case according to the reason for which the concept is invoked. Merely employing a concept of the state will not necessarily lead to any greater definitional clarity – for the concept begs just as many questions as the terms 'public' and 'private'.

Further analysis of the basis of judicial review can be found in the divergent approaches to Canadian law advocated in chapters 13 to 15. These chapters consider how interventionist the courts should be in terms of the *grounds* of judicial review in Canadian law, rather than focusing on the public law-private law divide. They might, in this respect, be felt to go beyond the stated remit of the collection (see Taggart's introduction), although this perhaps confirms the argument made earlier – in relation to Hunt's contribution – that any analysis of the basis for judicial review must address both the grounds *and* the scope of the courts' supervisory jurisdiction. The three chapters certainly show that debate about this topic is not confined to English law. In chapter 13, Dyzenhaus seeks to

 © The Modern Law Review Limited 1999

characterise modern defenders of the *ultra vires* model as positivist descendants of Bentham, who see the legislature as the sole source of legitimate legal authority. Supporters of rival theories are characterised as the antipositivist heirs of Blackstone, in that they see the common law as the guarantor of the rights and liberties which the courts are required to defend. The central question is therefore 'whether the legal limits of an agency's power are just those limits the legislature intended to set, or whether the common law sets limits as well' (p 284). Views as to how far the courts should defer to Parliament will vary between the two camps as well as among the theorists within each. Dyzenhaus rejects the notion that the courts should adopt what he describes as a posture of submissive deference – inherent in Dicey's theory of Sovereignty – whereby they simply submit to the will of the legislature. A preferable standpoint, based on democratic antipositivism, is one of deference as respect – whereby the courts pay respectful attention to the reasons offered for a decision by the legislature or the executive, but will nevertheless scrutinise such decisions for failure to respect an equality-driven conception of the rule of law. Deference as respect therefore requires that judges defer to an administrative determination 'on the basis of whether the agency has justified its determination in terms of its commitment to the value that provides the rationale for the administrative state' (p 306). This allows for judicial sensitivity to administrative bodies' own sense of how they can best respond to their mandates.

The main uncertainty in Dyzenhaus's argument is precisely *how* deferential the courts should be. It seems that deference as respect may be unable, on its own, to provide clear guidance on this point – something which is neatly encapsulated in the last footnote of the essay, where Dyzenhaus 'leave[s] dangling the question whether judges and tribunals should obey the legislature's commands when these are inequitable' (p307, n67). Dyzenhaus's reference to the need for judicial sensitivity to an agency's mandate is unlikely to help us here, given the many well-known problems of mandate-based political theories – problems which may indeed be exacerbated by privatisation and contracting out, which could be seen as creating a chain of mandates of varying content and strength. Dyzenhaus's analysis can be contrasted with the more traditional, Parliamentary intent-based approach advocated by L'Heureux-Dube in chapter 14. She suggests that administrative law is concerned to regulate powers delegated by the legislature to the administration. The courts must therefore adopt a pragmatic and functional approach (it should be noted that this is *not* the same approach as that developed by Craig in chapter 10, despite the similarity of labelling) when determining whether they are empowered to rule on a given issue, in order to prevent them from granting themselves jurisdiction over policy-related questions which the legislature has left to the tribunal to determine. Respect for the presumed intention of the legislature is crucial to this approach, and L'Heureux-Dube suggests that Canadian courts have shifted from an interventionist to a more deferential standard when judicially reviewing administrative tribunals for jurisdictional error. This is an interesting argument to advance given the Charter-based constitutional framework in Canada, but it is nonetheless subject to the standard questions which can be directed at *ultra vires* theories of judicial review in English law – namely, whether it is possible for courts meaningfully to discern the intention of the legislature, and whether this exercise is any event more than a mere 'fig leaf' hiding the otherwise naked exercise of judicial power.

Rather less satisfactory, conceptually, are some of the arguments made in chapter 15 by Harvison Young, who suggests that parallels may be drawn between administrative law and feminist theory when analysing the degree to which courts

© The Modern Law Review Limited 1999

should be deferential to administrative bodies. Harvison Young develops the idea of a functional or functionalist approach, but in a manner which is closer to Dyzenhaus's analysis than to L'Heuruex-Dube's. She suggests that such an approach will be concerned to take the purpose and nature of the administrative body under scrutiny seriously, according to the context in which that body operates. Ordinary courts should not be seen as superior to specialist decision-makers, and should adopt a posture of deference as respect towards the justifications offered by a decision-maker by being open and inclusive towards them, and sympathetic towards their perspective. According to Harvison Young, this approach might be characterised as feminist and pluralist in nature. The problem here, however, is that while chapter 15 contains some useful analysis of feminist legal thought, Harvison Young's argument effectively stands much of that thought on its head. One aspect of feminist theory is, as Harvison Young rightly argues, concerned with inclusive as opposed to confrontational modes of decision-making. But more fundamentally, feminism is concerned with the relations between those who have and those who lack power. The basic problem with talking about judicial 'sympathy' towards an administrative decision-maker is that both decision-makers and courts are, by definition, powerful (perhaps 'masculine') bodies. Talk of an inclusive approach is, in this context, rather more likely to allow a powerful body to get away with violating the rights of a powerless person – the very antithesis of a desirable solution from a feminist standpoint – rather than promoting any notion of real equality. From a feminist standpoint, inclusiveness and sympathy could quite rightly be required of the decision-maker or court in relation to the *citizen*, but surely not – other things being equal – as between the two institutions themselves.

The traditional distinction between public law and private law is discussed more broadly in three other essays. In chapter 8, McLean conducts a useful historical analysis of the *ultra vires*-driven 'source of power' test as a means of determining which bodies should be subject to judicial review, and suggests that it may now be more appropriate to focus on the nature of the function of the relevant body in a democratic system. Chapters 9 and 11 are, to an extent, in opposition to one another. In chapter 11, Oliver argues, via an analysis of English law, that five key values – autonomy, dignity, respect, status and security – are being developed by the courts in *both* public and private law, revealing the similarities between the two. Oliver concedes, however, that the freedom of action given to public bodies and private individuals will differ, and that higher order duties may properly be imposed on public bodies. In chapter 9, by contrast, Beerman openly acknowledges that there is and should be a substantive divide between the duties and responsibilities imposed on state bodies and private citizens or organisations in US law.

It has already been noted that the chapters dealing with the grounds of judicial review fall outside the collection's stated remit, although their relevance becomes clear when one considers the basis of the courts' supervisory jurisdiction. Such a reconciliation is not, however, possible in the case of chapter 12, by Janisch and Levi. This chapter provides a useful and informative case study of rule-making versus discretion in the policing context, and refines many of the arguments advanced by K.C. Davis concerning the comparative merits of discretion and rules (a point also taken up by Aronson in chapter 3). Considered in its own right, the chapter is innovative and worthwhile, but its role in this particular collection of essays remains unclear

Whether a public law-private law divide should continue to exist – and if so, where it should fall – can, ultimately, only be answered by arguments of political

 © The Modern Law Review Limited 1999

(and more particularly constitutional) theory. Challenges can certainly be levelled at many of the arguments developed in *The Province of Administrative Law*, but this should not detract from the conclusion that it is, on balance, a collection of extremely useful, very high quality work which pushes the debate about the public law-private law distinction a long way forward. Whatever one's position in this debate, the book deserves to be a standard point of reference for many years to come.

*Nicholas Bamforth**

Christopher Allen, **The Law of Evidence in Victorian England**, Cambridge: Cambridge University Press, 1997, xvi + 205pp, hb £35.00.

This book is a fine addition to the *Cambridge Studies in English Legal History* series. Based on Dr Allen's PhD thesis, 'Non-Benthamite Influences on the English Law of Evidence, 1828–1898' (University of London 1994), the book argues convincingly for a re-examination of Bentham's role in the reform of evidence law. The traditional view put forward by Maine in 1881, Dicey in 1914, Holdsworth in 1952, Hart in 1982, and Landsman in 1990 has placed Bentham at centre stage. Bentham and his followers, according to the traditional account, 'suggested and moulded the entire legislation' of the period (p 4, quoting Maine). His writings 'inspired the great statutory reforms' and had 'a profound impact on the development of the rules of evidence' (p 7, quoting Hart and Landsman). Not so, responds Dr Allen persuasively. A close examination of the debates in Parliament, the institution from which the statutory reforms emerged, reveals that Bentham's figure was, at best, remote. Bentham did have a role to play, Dr Allen observes, 'but not the only or even the main part' (p 182).

The book concentrates on the movement to lift restrictions on the competency of witnesses. At the beginning of the nineteenth century, potential witnesses were prohibited from testifying if they suffered from a 'defect of religious principle' (p 50), meaning that they either refused to take an oath or did not believe in the oath's sanctity; if they had committed a serious crime or one involving falsification; or if they had an interest in the case. By the end of the century, these grounds for disqualification had been abolished by statute. Taking each basis for incompetency in its turn, Dr Allen examines Bentham's relevant writings and then the history of the reforming legislation.

The book focuses first on incompetency from defect of religious principle. At common law, no witness could testify without taking an oath, a procedure viewed as a 'safeguard for the strict observance of truth' (p 51, quoting Phillipps' 1852 treatise on evidence). Potential witnesses who objected to oaths for religious reasons, Quakers for example, or those who held no belief in an afterlife, were not permitted to testify. Bentham criticised these prohibitions on multiple grounds: in his view, oaths were not effective at deterring untruthfulness; it was inconsistent to require oaths at trial but not in pleadings; people forbidden from testifying were denied access to justice; the refusal to hear testimony from potential witnesses reflected judicial laziness; and the insistence upon oaths furthered the 'sinister interests of the established Church' (pp 52–54). However, each of these arguments,

* Robinson College, Cambridge

© The Modern Law Review Limited 1999

as Dr Allen carefully demonstrates, played 'little or no part' in the debates leading to statutory reform (p 93). The reformers never sought the abolition of oaths, nor did they believe that oaths were simply an instrument of self-interested judges and bishops. Instead, they argued that the exclusion of relevant testimony hampered the law's ability to protect personal security and private property (pp 75–78). By allowing witnesses to affirm, rather than swear to, the veracity of their statements, the reformers maintained that more evidence would be available to prevent 'just debts [from being] avoided or property wrongly taken away' (p 78). 'It was entirely unimportant to the witness whether their testimony was received or not', the argument ran, 'but it was a serious misfortune to a tradesman suing a debtor if the only person who could prove the debt...entertained conscientious scruples against taking an oath' (pp 77–78). This position met initially with resistance, but in 1854 witnesses with conscientious objections to oaths were granted the right, in civil trials, to give evidence on affirmation (pp 59–60). In 1861, this right was extended to criminal trials, and in 1869 the requirement of conscientious objection was removed, thereby enabling testimony on affirmation from atheists and other nonbelievers (pp 60–61).

Dr Allen then turns to incompetency as a result of prior crimes, known as 'incompetency from infamy' (p 95), and incompetency caused by having an interest in the case. At common law, testimony was not permitted from persons previously convicted of crimes 'inconsistent with the common principles of honesty' (p 95, quoting Starkie's 1833 treatise on evidence), nor was evidence allowed from anyone with an interest in the outcome of the proceedings, a category that pointedly included the parties. Bentham attacked these prohibitions, arguing that it was better to admit potentially false evidence than to exclude it. In his view, the law had 'abundant remedies' against being misled by lies, including the 'natural sagacity of the jury' and the 'cultivated sagacity of the judge', but the exclusion of evidence was 'disastrous' because it led inevitably to wrong decisions (p 99, quoting the *Rationale on Judicial Evidence*). Bentham was particularly firm about the need for evidence from the litigants themselves: 'hear everybody, but most attentively of all, and first of all, those who are likely to know most about it – the parties' (p 99, again quoting the *Rationale*). These opinions had only a limited influence on the reformers, Dr Allen argues sensibly, because none of the initial advocates for change was pressing for the competency of the parties. When statutory reform occurred in 1843, the legislation eliminated incompetency from infamy and *non-party* incompetency from interest; but the litigants remained ineligible to testify. It was not until 1851 that parties in civil cases were permitted to speak, by which time many of the arguments for reform focused on facts that had not attracted Bentham's attention, such as the regular use of party testimony in commercial arbitrations and in courts other than the superior courts of common law. The impetus behind the 1851 reform may have been partly Benthamite, Dr Allen observes cautiously, but it is more likely that the reforms 'developed naturally from existing institutions and earlier precedent: the competency of parties in [other] tribunals ... and the inroad into incompetency from interest already made in 1843' (p 122).

The book's final chapter (not counting the conclusion) focuses on the slow process by which competency was extended to criminal defendants. Here too, Dr Allen persuasively demonstrates that Bentham's influence was 'indirect and slight' (p 175). At common law, persons accused of crimes were not permitted to speak upon oath, a practice Bentham derided as 'groundless and utterly indefensible' because it kept out the best and most direct evidence – and in so doing, encouraged

© The Modern Law Review Limited 1999

the use of less reliable proof, such as hearsay (p 129, quoting the *Rationale*). Defendants should not only be allowed to testify, Bentham argued, but required to do so or have adverse inferences drawn from their silence. As he put it, 'all suspected persons who are not guilty, court [the opportunity to speak]; none but the guilty shrink from it' (p 128, again quoting the *Rationale*). After several abortive attempts, reform came in 1898 with the passage of the Criminal Evidence Act, which extended the right to testify to all accused persons. Three arguments for reform, explains Dr Allen, dominated the debates: first, that improved methods of discovering the truth were needed to convict guilty defendants and protect innocent ones; second, that defendants were disadvantaged because their opponents – prosecutors and police – were permitted to speak on oath; and third, that the earlier extension of competency to parties in civil cases rendered the accused's incompetency an indefensible anomaly (pp 161–167). In the end, these arguments and others carried the day. But as Dr Allen demonstrates, the compulsory examinations viewed as essential by Bentham were never seriously contemplated by the reformers, nor did they require negative inferences to be drawn from a defendant's decision to remain silent (pp 158, 175). For these reasons, he reasonably concludes that Bentham's influence on the 1898 legislation was limited at best.

As I indicated at the beginning of the review, this book is an example of good legal-historical research. Dr Allen has marshalled an impressive array of primary sources, and his interpretations and arguments are lucid and well-presented. No book is perfect, however, and one relatively minor quibble is worth mentioning. Two important articles, both of which appeared in the summer of 1996, are noticeably absent from the footnotes and bibliography: Professor Langbein's work on mid-eighteenth-century evidence law (96 *Columbia Law Review* 1168, June 1996) and Professor Alschuler's historical treatment of the right to remain silent (94 *Michigan Law Review* 2625, August 1996). Dr Allen's book was published more than year later, in September 1997, and thus I was left wishing that these articles had been incorporated into the book's argument, or at least mentioned. But this is probably a counsel of perfection, and it would be an injustice to this fine book to end the review on a discordant note. Let me therefore conclude by returning to my main point: this is an excellent book, well-researched and well-argued, that convincingly re-evaluates Bentham's role in the nineteenth-century reform of evidence law. It is a welcome addition to the scholarly literature.

*T.P. Gallanis**

Lilian Edwards and Charlotte Waelde (eds), **Law & the Internet**, Oxford: Hart Publishing, 1997, xix + 264pp, pb £25.00.

This book is a timely contribution to a topical subject and addresses a question currently vexing this reviewer: namely, does the Internet create a need for new legal frameworks? The book concentrates on intellectual property, electronic commerce and liability for Web site content. This reviewer was disappointed to find only incidental discussion of the privacy implications of the Internet.

* College of Law, Ohio State University.

© The Modern Law Review Limited 1999

Authorship in the collection is split between practitioners and academics, which makes for interesting contrasts of style and content.

In terms of intellectual property, there is discussion in two areas. First, ownership of domain names, which does not seem to create an immediate need for new legal rules. Secondly, copyright more generally is discussed. This has generated significant legislative activity at international level, in particular the World Intellectual Property Organisation Copyright Treaty 1996. With regard to electronic commerce, it seems that the Internet generates a need for adaptation of existing rules as well as a need for some altogether new rules. Lars Davies gives an intricate account of computer communication in order to apply thereto contractual rules of offer and acceptance. Hector MacQueen raises the thorny issue of whether software is a good or a service and discusses some terms commonly found in software licences. The moulding of existing rules to new circumstances only provides a partial solution in the contexts of tax and contract law. The phenomenon of digital signature has generated significant legislative activity (although not yet in England). The tax implications of the Internet seem to overshadow other commercial complications, since pinpointing and policing Internet transactions is so difficult. Sandra Eden addresses these implications and sees a growing need for new rules.

This book will be stimulating to those interested in the legal implications of the Internet. This reviewer would have preferred the book to have used more extensively the question posed above as a strand to link disparate topics together, but that preference may well not be shared by all. Also, to have included authors based outside the UK might have served the book well. The presentation of the book is faultless.

*John Dickie**

G.P. Stapledon, **Institutional Shareholders and Corporate Governance**, Oxford: Clarendon Press, 1996, xxxii + 376 pp, hb £55.00.

Since the publication of Berle and Means's *The Modern Corporation and Private Property* in 1932, the 'separation of ownership and control' in listed public companies (ie, companies whose shares are listed and therefore can be traded on a stock exchange) has been regarded as the central problem in corporate governance. Shareholders in such a company are perceived to be unable and unwilling to behave like the owners which they supposedly are and to ensure that management delivers the maximum return possible. Shareholder weakness appears to give management some leeway to use a company's resources for their own benefit, for example by awarding themselves excessive remuneration or embarking on an unnecessary expansion of the company's activities. In economic terms, this is likely to be inefficient and lead to a misallocation of resources. The challenge in corporate governance has been to reduce this problem in a way that is cost-effective and does not undermine the efficiency gains of separating the functions of management and shareholding. Thus, one way of sharpening the interest of shareholders in their company's affairs would be to remove their limited liability, but such a move would not be welcomed by shareholders and unlikely to ensure an adequate supply of equity capital. Some economic analysts of the company have

* School of Law, University of Warwick.

© The Modern Law Review Limited 1999

argued that there is not really a problem at all because management's scope for serving their own interests is effectively constrained by competition in the markets in which they operate, including the market for equity capital and its concomitants, share prices and the threat of takeover or 'the market in corporate control'. A crucial aspect of this approach is to view a company's shareholders not as owners in the traditional sense, but rather as parties to a complex contractual arrangement concerning the provision of equity capital.

The nature of the typical shareholder in a public company has been a key feature in debates about corporate governance. The classic exposition of the separation of ownership and control presented such shareholders as having a small stake and thus a small interest in the company and its management, a syndrome exacerbated by their limited liability and the common practice of countering the risk in any one shareholding by holding a diversified portfolio of investments. Such shareholders would face major collective action and co-ordination problems if they wished to respond actively to poor or inefficient management in a company rather than just exit the company by selling their holding and investing elsewhere. They were thus seen as a weak and ineffective counterweight to the concentrated powers of management. However, the growth of institutional investment has substantially changed this picture. Instead of public companies having a mass of small and insignificant shareholders, there now appear to be powerful players on the ownership side of the governance equation who have become the focus of great expectations. This transformation has been noted with interest in the recent reports on corporate governance sponsored by a group of concerned organisations including the London Stock Exchange and the Confederation of British Industry. Thus, the *Cadbury Report*, published in December 1992, observed that the 'proportion of shares held by individuals and by institutions has broadly reversed over the past thirty years so that institutional shareholders now own the majority of shares of quoted companies' (para 6.9). The ensuing *Hampel Report*, published in January 1998, which was commissioned to review the implementation of *Cadbury's* recommendations, said that '60% of shares in listed UK companies are held by UK institutions – pension funds, insurance companies, unit trusts and investment trusts' and that of the remaining 40 per cent 'about half are owned by individuals and half by overseas owners, mainly institutions' (para 5.1). From this, *Hampel* concluded that 'a discussion of the role of shareholders in corporate governance will mainly concern the institutions, particularly UK institutions'. Both reports looked to these institutions to use their influence to ensure that listed public companies complied with their respective codes of conduct and directed specific recommendations at them.

Stapledon's book is therefore a welcome and valuable contribution to the corporate governance debate and provides a wealth of information and background material about institutional shareholding in the United Kingdom and Australia, drawing on extensive empirical research conducted by the author. This work examines the growth of financial institutions in both countries, the role which these organisations already play in corporate governance and what might be expected of them in the future. The author uses Australia as a point of comparison because a much larger proportion of its corporate sector consists of subsidiaries of foreign multinationals and companies controlled by dominant shareholders and therefore raises different governance problems. He depicts the United Kingdom's system of governance as an 'outsider' system, which emphasises the fact that very few listed public companies have shareholdings large enough to exert a direct controlling influence on the company's management structure. Instead, ultimate control is

© The Modern Law Review Limited 1999

dispersed into small securities held by outsiders. In Australia, however, with far more companies subject to dominant shareholders, there is in fact an 'insider' system running in parallel, similar to the ones associated with those more familiar points of comparison, Germany and Japan. The book completes its task in four parts. The first one provides a framework for the subsequent analysis. The second and third parts provide detailed examinations of institutional shareholders and their current role in corporate governance in the United Kingdom and Australia respectively. The fourth part looks to the future, considering whether the institutions role should be extended, whether it can be and what regulatory reforms would be appropriate.

In Part I, the author describes the governance problem facing shareholders in a listed public company in some detail. This problem is presented in terms of 'agency costs', that is the cost to shareholders of any deviation by management from their proper role of maximising the shareholders' return plus the cost of the devices used to reduce the scope for deviation. He regards the monitoring of a company's management by its shareholders as one element in the environment of constraints and incentives which have been identified as reducing agency costs, along with legislation, stock market regulation, market forces, directors' duties and profit-related remuneration. He argues that shareholder monitoring is important because of serious weaknesses in the other elements. Thus, he does not accept that the market in corporate control and other market forces are adequate constraints by themselves in the way some economic analysts have claimed and notes that problems of enforcement undermine the value of directors' duties. The growth of institutional shareholding is therefore a significant development because it appears to have increased the likely effectiveness of shareholder monitoring and lessened the separation of ownership and control.

In Part II, the author focuses on institutional shareholders in the United Kingdom, describing the various financial institutions involved, how their funds are actually managed and the collective action bodies which represent their interests. This exposition reveals the complexity which underlies references to 'institutional shareholders' and the fact that these organisations have governance problems of their own since the management of their investments is often delegated to external fund managers. This can hinder or complicate active monitoring of the companies in which they hold shares. The author then analyses the role which institutional shareholders have actually played in corporate governance, providing detailed surveys of the issues which have attracted their attention and of the methods which they have used to express their concern. He notes that their interest has generally been limited to matters affecting the value of shares as an investment and that they have taken scant interest in the wider ethical and political issues which have sometimes featured in debates on corporate governance. The institutions have tended to monitor behind the scenes through routine meetings and other kinds of dialogue between fund managers and management. They have avoided litigation to deal with mismanagement and have been reluctant to use their voting rights, which is a consequence of their own governance problems since voting entails active decision-making. The author has nevertheless uncovered some recent developments of interest which reflect the institutions' greater importance as shareholders. Pressure from small coalitions of institutional shareholders has sometimes contributed to changes of management in problem companies. Also, institutions have increased to their links with non-executive directors (NEDs), although this is only a slight increase from a very low base. NEDs are viewed as an important potential constraint on corporate

© The Modern Law Review Limited 1999

management and, like the institutions themselves, featured prominently in the recommendations of the *Cadbury* and *Hampel* Reports. However, as the author notes, their effectiveness in practice is questionable, not least because they are likely to identify with the interests of management. The author identifies several reasons for this, including management's instrumental role in their selection and the fact that they are often executive directors elsewhere, ensuring that management has their sympathy and empathy. One major weakness in their role is that, whilst notionally independent of management, they are not dependent on any alternative source of power. One response to this weakness would be for the institutions to nominate NEDs themselves, but the author notes that they have rarely done this. This is in marked contrast to their approach when investing in non-listed companies. The likely reasons for this include their reluctance to become insiders, the need for NEDs to represent all shareholders and not just a particular nominee and the time and resources that would be entailed in adopting this practice generally. The author does not see any prospect of change in this respect.

In Part III, the author adopts the same approach in analysing the governance role of institutions in Australia. Again, it is fund managers who would have to engage in monitoring. Their involvement is similar to that in the United Kingdom, but has a shorter history and has been narrower in scope. One striking difference is that the boards of most large and medium-sized listed companies in Australia have for a long time had a majority of NEDs. However, in practice, like NEDs in the United Kingdom, they tend to be close to management and this difference has had little impact in practice. The author also notes the difficulties facing the institutions, along with other minority shareholders, in companies which are controlled by a dominant shareholder or clique. These difficulties stem from the fact that management is immune from any takeover against the wishes of the dominant shareholders, which skews the market in corporate control and undermines a key component in the array of constraints usually faced by management. Since minority shareholders in a scenario of this kind do not pose a serious threat to management, regardless of their relative size and whether or not they act collectively, the value of monitoring is severely reduced.

In Part IV, the author concludes his analysis by considering how monitoring by institutional shareholders could be improved and whether this would lead to an improvement in corporate governance in the future. He advocates a greater level of institutional monitoring and argues that the advantages of this will outweigh the disadvantages. However, he also contends that this is unlikely to result from mere exhortation and therefore recommends some modest changes in legislation and stock exchange regulation in each country to facilitate this. He identifies the main advantages as reduced agency costs, since monitoring provides a more effective 'trip wire' than the market in corporate control, and also a greater emphasis on the long term by management. The latter benefit is at first glance ironic. Whilst 'short-termism' has been recognised as a problem facing economies with outsider systems of corporate governance, it has usually been attributed to the pressure on management to meet shareholder expectations of a high return. The growing influence of financially-driven institutional shareholders, in tandem with the market in control, has accordingly been regarded as increasing management's short-termism. The author therefore examines this problem in some detail to support his point of view. He argues that in part the problem stems from a misconception on the part of management as to the real expectations of shareholders and that a greater level of monitoring would reduce the scope for

© The Modern Law Review Limited 1999

this occurring. He rejects the idea that short-termism reflects a defect in the capital markets, which greater shareholder monitoring could intensify, and instead regards it as an inherent problem resulting from the fact that shareholders in outsider systems tend to be disinterested investors rather than committed to a company in which they hold shares. The impact of this on management is negative in that it does not encourage them to take a short term view, but gives them no incentive to do otherwise. From this perspective, the problem of short-termism is likely to persist, but would not be exacerbated by an increased level of shareholder monitoring.

The author advocates a modest package of reforms. They include proposals designed to remove some barriers to voting by institutions, especially in Australia, although he rejects the idea of compulsory voting. In this respect, his conclusions are echoed in the recommendations of the *Hampel Report*. He also recommends an alteration in the Stock Exchange rules so that the thresholds used to specify transactions which must be referred to the shareholders for approval are reduced from 25 per cent to 20 per cent. Despite the prevalence of companies with dominant shareholders in Australia, the author directs his recommendations at the outsider system, although he argues that monitoring by institutions who are part of an entrenched minority may still be important in combatting prejudice to minorities of this kind. However, it is arguable that the additional danger faced by shareholders in companies like this is part of the risk of shareholding as an investment and should be reflected in the price paid for the shares and their subsequent trading price. If this danger were excessive, companies of this kind would presumably face difficulty in raising equity capital and market pressures should lead to corrective safeguards. In the opinion of this reviewer, analysis of corporate governance does sometimes over-emphasise the apparent vulnerability of shareholders and overlooks the significance of other stakeholders. Excessive managerial remuneration, for example, may be an indication of a company's excessive power in the market place rather than of management's ability and willingness to exploit shareholder weakness.

This book is recommended reading for anyone who wishes to understand the nature of institutional shareholders, how they deal with their investments in public companies and the prospects for relying on them to improve the quality of corporate governance in the future.

*Andrew Griffiths**

* Faculty of Law, University of Manchester.

© The Modern Law Review Limited 1999

Information for Contributors

Contributions to and correspondence concerning the various sections of the Review should be sent to the relevant Editor:

Articles: Hugh Collins, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Legislation and Reports: R.C. Simpson, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Case Notes: Roger Brownsword, Faculty of Law, The University, Crookesmoor Building, Conduit Road, Sheffield S10 1FL.

Books for Review, Review Articles and Reviews: Neil Duxbury, Faculty of Law, University of Manchester, Oxford Road, Manchester M13 9PL.

General Correspondence: Tim Murphy, Law Department, London School of Economics, Houghton Street, London WC2A 2AE.

Section Editors will be happy to discuss drafts and proposed contributions with contributors. All contributions should be typed, double spaced and include an estimate of the number of words, including footnotes. Contributions should use footnotes which make cross-references within the text as sparingly as possible. Contributions must conform to the Review's house style; a style sheet will be provided by the Production Editor on request.

Articles do not usually exceed 12,000 words (excluding footnotes). Longer or shorter articles can be considered.

Subscription Information 1999

The Modern Law Review is published six times a year. Subscription rates are as follows:

Annual subscription for unbound parts only	
UK and Europe	£73.00
Rest of the world	£75.00
North America	US$124.00*
UK and European law teachers and law students	£21.00
Commonwealth law teachers and law students	£26.00/US$43.00*
Annual subscriptions for Bound Volume Service: unbound parts plus a bound volume	
UK and Europe	£121.00
Rest of the world	£128.00
North America	US$211.00*
UK and European law teachers and students	£59.00
Commonwealth law teachers and law students	£62.00/US$102.00*

Back Issues

Single issues of the unbound parts of the last three volumes are available from the publishers at the current price for unbound parts:

	£15.00 (£15.00 for ROW, US$25.00 in North America)*
Single bound volumes	£61.00 (£63.00 for ROW, US$104.00 in North America)*
Complete sets of bound volumes	£2071.00 (US$3417.00 in North America)*

*Canadian customers/residents please add 7% GST.

Orders, sample copy requests, claims and other subscription enquiries should be addressed to: Blackwell Publishers Journals, PO Box 805, 108 Cowley Road, Oxford OX4 1FH, UK. All subscriptions are supplied on a calendar year basis (January to December).

US Mailing

Periodicals postage paid at Rahway, New Jersey.
Postmaster: send address corrections to *The Modern Law Review*, c/o Mercury Airfreight International Ltd Inc., 365 Blair Road, Avenel, NJ 07001, USA (US mailing agent).

Microform

The journal is available on microfilm (16 mm or 35 mm) or 105 mm microfiche from the Serials Acquisitions Department, University Microfilms Inc, 300 North Zeeb Road, Ann Arbor, MI 48106, USA.